



UNIVERSITAT DE
BARCELONA

Generative Deep Learning for Cancer Image analysis

Richard Osuala

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

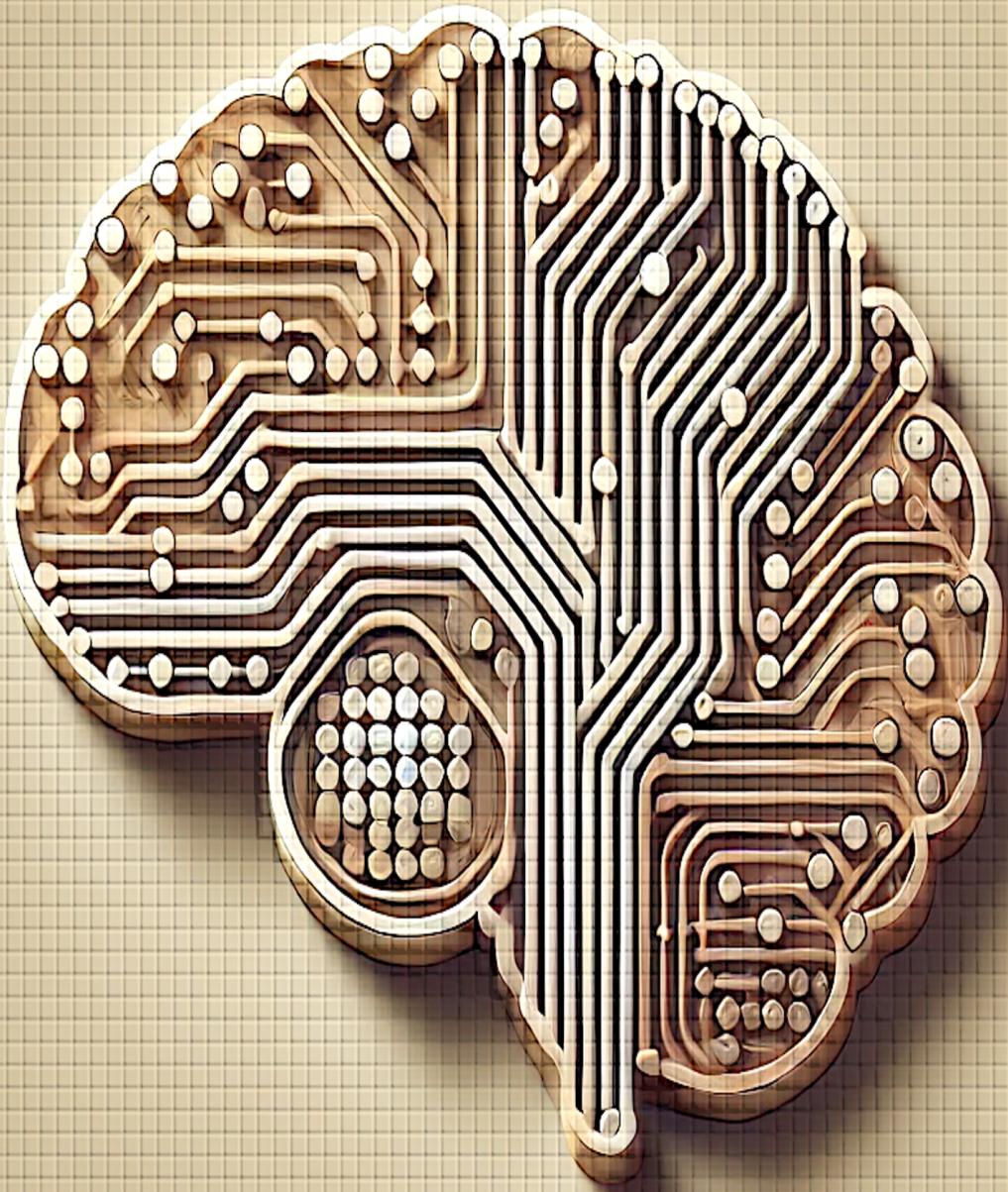
ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



UNIVERSITAT DE
BARCELONA

GENERATIVE DEEP LEARNING FOR CANCER IMAGE ANALYSIS



Doctoral Thesis

Richard Osuala



UNIVERSITAT DE
BARCELONA

Doctoral Program on Mathematics and Computer Science

GENERATIVE DEEP LEARNING FOR CANCER IMAGE ANALYSIS

This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy
(PhD) at the Universitat de Barcelona.

Author: Richard Obinna Osuala

Supervision

Director: Dr. Karim Lekadir

Co-Director: Dr. Oliver Díaz

May 2025



UNIVERSITAT DE
BARCELONA

Doctorat en Matemàtiques i Informàtica

APRENENTATGE PROFUND GENERATIU PER A L'ANÀLISI D'IMATGES DEL CÀNCER

Aquesta tesi es presenta com a requisit parcial per a l'obtenció del grau de Doctor en
Filosofia (PhD) a la Universitat de Barcelona.

Autor: Richard Obinna Osuala

Supervisió

Director: Dr. Karim Lekadir

Co-Director: Dr. Oliver Díaz

Maig 2025



UNIVERSITAT DE
BARCELONA

Doctorado en Matemáticas e Informática

APRENDIZAJE PROFUNDO GENERATIVO PARA EL ANÁLISIS DE IMÁGENES DEL CÁNCER

Esta tesis se presenta como requisito parcial para la obtención del grado de Doctor en
Filosofía (PhD) en la Universitat de Barcelona.

Autor: Richard Obinna Osuala

Supervisión

Director: Dr. Karim Lekadir

Co-Director: Dr. Oliver Díaz

Mayo 2025

Declaration of Authorship

I, Richard Obinna OSUALA, declare that this thesis titled, "Generative Deep Learning for Cancer Image Analysis" and the work presented in it are my own. I confirm that:

- This work was done while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is clearly attributed.
- GPT4o was used as tool to proofread parts of the thesis and to generate part of the cover image. Where I have quoted from the work of others, the source is given. With the exception of such quotations and GPT4o reformulations, this thesis is entirely my own work.

I told you everything I know. From this point forth, we shall be leaving the firm foundation of fact and journeying together through the murky marshes of memory into thickets of wildest guesswork.

Albus Dumbledore

If the wonders of life unfold in a Markovian chain of events, then the completion of this thesis chaining the author long remained a particularly unlikely event. Despite causing a significant increase of entropy in the author's life, hope is left it may do the contrary for its readers' minds.

The Author

Acknowledgements

This PhD journey has been a whirlwind of learning, discovery, and growth, both personal and academic. It would not have been possible without the support, generosity, and presence of many people, each of whom has contributed in their own meaningful way.

I am deeply grateful to my supervisors, Dr. Oliver Díaz and Dr. Karim Lekadir. Their mentorship, vision, support, and trust were foundational in shaping both the course of this research and my growth as a scientist. Karim and Oliver taught me how to focus on what is important, to continuously progress and remain open to new endeavours, to distill complex ideas into clear and ambitious objectives, and to develop and drive my own research agenda. I feel fortunate to have contributed to the development of the Barcelona Artificial Intelligence in Medicine (BCN-AIM) lab under their guidance, which has been an enriching and formative experience. I would also like to express my heartfelt thanks to Dr. Julia Schnabel, as well as her team and her assistant, Sandra Mayer, who hosted me during my research stay at Helmholtz Munich and the Technical University of Munich. Julia's openness, insight, and encouragement greatly broadened the scope of my research and deeply enriched my academic experience.

I am deeply grateful to my colleagues at BCN-AIM at the Universitat de Barcelona, and to my colleagues at Helmholtz Munich and the Technical University of Munich. Sharing this journey with you, through challenges, insightful discussions, memorable moments, and even fun activities outside of work, has been one of the most rewarding aspects of my PhD. I would like to especially thank Kaisar, whose technical expertise, mentorship, and generous guidance taught me not only how to conduct research, but how to shape ideas into meaningful scientific contributions. To Lidia and Smriti, thank you for walking this path with me as fantastic teammates in the Oncology group. And to Daniel, thank you for our enjoyable collaboration, and for the encouragement

and sharp critical thinking you brought to it. To all BCN-AIM and IML colleagues: Your friendship, insights, and solidarity made the research environment not only productive, but vibrant and meaningful. Working with you has been both personally and intellectually enriching, and I carry those memories forward with real gratitude.

I also wish to thank all academic partners and further collaborators who contributed to this journey and to my research. To name just a few - Grzegorz, Zuzanna, Bennet, Apostolia, Preeti, Cosmin, Stefan, Anneliese, Socayna, Noussair, Nick, Victor, Carlos, Eloy, Akis, Walter - your expertise and engagement elevated this work, and I am truly grateful for the opportunities to learn from you. If I have unintentionally left anyone out, please know that your contributions are equally appreciated. Moreover, I am grateful for the EuCanImage Horizon 2020 Project (Grant Agreement ID: 952103), which provided the financial support for my PhD position, an opportunity for which I remain deeply appreciative. To the EuCanImage project managers over the years, Isabell, Anais, and Jamilia, and also the BCN-AIM management team including Katharina, Paloma, Xènia, Cathy, and Cecilia, thank you for your exceptional coordination, support, and professionalism. Your efforts played a crucial role in keeping the projects, and by extension my work, running smoothly.

I am grateful to the many teachers, mentors, friends, colleagues, and kind-hearted people I have encountered from school through university and beyond, who generously shared their knowledge, took the time to explain, guide, and inspire, and helped expand my understanding of the world. Your quiet acts of support and encouragement have made a lasting impact on my life and learning. Special thanks to Dr. Ognjen Arandjelović, whose mentorship during my Master's thesis first lit the spark that led me down this path. His belief in my potential, along with his insights and encouragement, played a pivotal role in setting me on the path to research. I also wish to thank former colleagues, whose stories about their own PhD journeys helped plant the seed that this path was not only possible, but worth pursuing.

To my family, thank you for your love, support, resilience, and belief in me through every step of this long journey. To my parents, Eva and Eddie, you have been the roots from which everything else has grown, and this will always remain part of the story we have written together. To my friends, thank you for grounding me and reminding me of life beyond deadlines, papers, and experiments. Your presence brought joy and balance when I needed it most. I want to thank my partner, Marie, for being by my side during the final stages of this journey, you gave me the strength, love, perspective, and calm to keep going and to see it through. I also thank my former partner, Carme, for being a pillar encouraging and supporting me, for starting our PhD journey together with love

and courage guiding us on our way into uncertain terrains. I would also like to thank her parents Carmen and Lluís for all their warm kindness and support throughout these years.

Looking back, I realize that a PhD is not only a technical achievement, but a deeply personal evolution. It is an exercise in perseverance, humility, and self-discovery. The knowledge gained is not only in the research itself, but in learning to navigate complexity, failure, and the slow, quiet triumphs of understanding. For all of this, I am grateful. This thesis is dedicated to those who walked this path with me, and to the belief that curiosity, kindness, and collaboration can take us farther than we imagine.

Summary

Accurate and timely detection of cancerous lesions in medical imaging is essential for effective treatment. However, the diagnosis remains challenging due to, among others, tumor heterogeneity, imaging constraints, and observer variability. Deep learning architectures, such as convolutional and transformer-based networks, have been showing promise in improving cancer image analysis by learning complex patterns within features from raw imaging data, allowing for earlier, more precise detection and consistent, data-driven decision-making. Despite its potential, clinical adoption of deep learning is restricted by the need for large, annotated training datasets, which are scarce due to privacy and labeling cost constraints, as well as due to its variability in performance when applied in settings of domain shift, imaging artifacts, variation in imaging protocols, and patient populations.

This thesis identifies and addresses the challenges in deep learning for cancer imaging through five core publications that propose novel frameworks, methods, and solutions. First, a large-scale survey of generative models in cancer imaging is conducted leading to the identification of key challenges and problems in the field alongside ideation of potential solutions. Based on these findings, the SynTRUST meta-analysis framework is derived to assess the trustworthiness and maturity level of cancer image synthesis studies and solutions. Second, conditional generative adversarial networks (GANs) are applied to the challenges of scarcity of cancer images and tumor annotations, by simulating dynamic contrast-enhanced breast magnetic resonance imaging (DCE-MRI) sequences. Without relying on physical contrast agents or DCE-MRI data from real patients, this approach enables unsupervised tumor detection, localization, and characterization, along with providing synthetic training data for increasing the robustness of downstream task models such as tumor segmentation models. Third, a multi-conditional latent diffusion models is developed to translate non-contrast enhanced MRI images into variable time-dependent synthetic DCE-MRI images localizing tumors

and predicting their contrast enhancement kinetic patterns. Addressing the need for respective quantitative evaluation metrics, the Fréchet Radiomics Distance (FRD) is proposed to measure (synthetic) image quality based on biomarker variability. Fourth, a mass malignancy-conditioned generative adversarial network (MCGAN) is proposed to generate synthetic data as privacy-preservation mechanism for training deep learning models without individual patient data. Via comparison and combination with differential privacy, the synthetic mammography data is shown to improve the performance in multiple privacy-preserving cancer classification scenarios. Fifth, the medigan library is introduced as sharing platform for pretrained generative models that enable researchers to generate high-quality synthetic data across diverse imaging modalities without requiring direct access to sensitive patient data. Additionally, medigan's generative models are comprehensively analyzed based on the Fréchet Inception Distance (FID) using both radiology domain-specific and standard domain-invariant feature extractors.

In conclusion, this thesis highlights the potential of generative deep learning to address the key challenges in cancer imaging, presenting novel methods in contrast enhancement simulation, tumor localization, privacy-preserving cancer classification, and tools for sharing and assessing generative models, paving the way for these models towards integration into clinical practice to the end of advancing healthcare for both individual patients and society at large.

Keywords: Deep Generative Models, Medical Image Analysis, Data Synthesis, Contrast Enhancement Simulation, Privacy-Preserving Machine Learning, Synthetic Data Evaluation.

Resum

La detecció precisa i oportuna de lesions canceroses en les imatges mèdiques és essencial per a un tractament eficaç. No obstant això, el diagnòstic continua presentant desafiaments a causa, entre altres factors, de l'heterogeneïtat tumoral, les limitacions de les imatges i la variabilitat entre observadors. Les arquitectures d'aprenentatge profund, com les xarxes convolucionals i les xarxes basades en transformers, han demostrat ser prometedores per a millorar l'anàlisi d'imatges de càncer mitjançant l'aprenentatge de patrons complexos dins de les característiques de les dades d'imatges mèdiques, cosa que permet una detecció més precoç i precisa i una presa de decisions coherent i basada en dades. Tot i el seu potencial, l'adopció clínica de l'aprenentatge profund està restringida per la necessitat de grans conjunts de dades d'entrenament anotades, que són escassos a causa de les restriccions de privacitat i del cost d'etiquetatge, així com de la variabilitat en el seu rendiment quan s'aplica en entorns amb canvi de domini, amb artefactes d'imatge, variacions en els protocols d'imatge i en diferents poblacions de pacients.

Aquesta tesi identifica i aborda els reptes de l'aprenentatge profund per a l'obtenció d'imatges de càncer a través de cinc publicacions principals que proposen marcs, mètodes i solucions innovadors. En primer lloc, es duu a terme un estudi a gran escala dels models generatius en la imatge del càncer que condueix a la identificació de desafiaments i problemes clau en el camp, juntament amb la proposta de possibles solucions. A partir d'aquests descobriments, es deriva el marc de metaanàlisi SynTRUST per avaluar la fiabilitat i el nivell de maduresa dels estudis i solucions de síntesi d'imatges de càncer. En segon lloc, s'apliquen xarxes generatives adversàries condicionals (GANs) als reptes de l'escassetat d'imatges de càncer i anotacions tumorals, simulant seqüències de ressonància magnètica de mama amb contrast dinàmic (DCE-MRI). Sense dependre d'agents de contrast físics ni de dades de DCE-MRI de pacients reals, aquest enfocament permet la detecció, localització i caracterització no supervisada de tumors, a més de

proporcionar dades sintètiques d'entrenament per augmentar la robustesa dels models per a tasques posteriors, com la segmentació de tumors. En tercer lloc, es desenvolupa un model de difusió latent multicondicional per transformar imatges de MRI sense realç de contrast en imatges sintètiques de DCE-MRI variables en el temps que localitzen els tumors i prediuen els seus patrons cinètics de realç de contrast. Per abordar la necessitat d'una mètrica quantitativa respectiva d'avaluació, es proposa la Distància Radiòmica de Fréchet (FRD) per mesurar la qualitat de la imatge (sintètica) basada en la variabilitat de biomarcadors. En quart lloc, es proposa una xarxa generativa adversària condicionada a la malignitat de masses (MCGAN) per generar dades sintètiques com a mecanisme de preservació de la privadesa per a l'entrenament de models d'aprenentatge profund sense dades de pacients individuals. Mitjançant la comparació i combinació amb la privadesa diferencial, es demostra que les dades sintètiques de mamografia milloren el rendiment en múltiples escenaris de classificació de càncer amb preservació de la privadesa. En cinquè lloc, es presenta la biblioteca medigan com a plataforma compartida per a models generatius preentrenats que permet als investigadors generar dades sintètiques d'alta qualitat en diverses modalitats d'imatge sense necessitat d'accedir directament a dades sensibles de pacients. A més, els models generatius de medigan s'analitzen exhaustivament basant-se en la Fréchet inception distance (FID) utilitzant tant extractors de característiques específics del domini radiològic com extractors de característiques estàndard invariants del domini.

En conclusió, aquesta tesi demostra el potencial de l'aprenentatge profund generatiu per afrontar els desafiaments clau en les imatges de càncer, presentant mètodes innovadors en la simulació de millora del contrast, localització de tumors, classificació de càncer preservant la privadesa i eines per compartir i avaluar models generatius, obrint el camí per a la integració d'aquests models en la pràctica clínica i el progrés de l'atenció sanitària, tant per als pacients individuals com per a la societat en general.

Paraules clau: Models Generatius Profunds, Anàlisi d'Imatges Mèdiques, Síntesi de Dades, Simulació de Millora de Contrast, Aprenentatge Automàtic amb Preservació de la Privacitat, Avaluació de Dades Sintètiques.

Resumen

La detección precisa y oportuna de lesiones cancerosas en las imágenes médicas es esencial para un tratamiento eficaz. Sin embargo, el diagnóstico sigue presentando desafíos debido, entre otras cosas, a la heterogeneidad del tumor, las limitaciones de las imágenes y la variabilidad entre observadores. Las arquitecturas de aprendizaje profundo, como las redes convolucionales y las basadas en transformers, han demostrado ser prometedoras para mejorar el análisis de imágenes de cáncer mediante el aprendizaje de patrones complejos dentro de las características de los datos de imágenes, lo que permite una detección más temprana y precisa y una toma de decisiones coherente y basada en datos. A pesar de su potencial, la adopción clínica del aprendizaje profundo está restringida por la necesidad de grandes conjuntos de datos de entrenamiento anotados, que son escasos debido a las restricciones de privacidad y costo de etiquetado, así como debido a la variabilidad en su rendimiento cuando se aplica en entornos de cambio de dominio, con artefactos de imagen, variaciones en los protocolos de imagen y en diferentes poblaciones de pacientes.

Esta tesis identifica y aborda los retos del aprendizaje profundo para la obtención de imágenes de cáncer a través de cinco publicaciones principales que proponen marcos, métodos y soluciones novedosos. En primer lugar, se lleva a cabo un estudio a gran escala de los modelos generativos en la imagenología de cáncer que conduce a la identificación de desafíos y problemas clave en el campo, junto con la propuesta de posibles soluciones. A partir de estos hallazgos, se deriva el marco de metaanálisis SynTRUST para evaluar la fiabilidad y el nivel de madurez de los estudios y soluciones de síntesis de imágenes de cáncer. En segundo lugar, se aplican redes generativas adversariales condicionales (GANs) a los retos de la escasez de imágenes de cáncer y anotaciones tumorales, simulando secuencias de resonancia magnética de mama con contraste dinámico (DCE-MRI). Sin depender de agentes de contraste físicos ni de datos de DCE-MRI provenientes de pacientes reales, este enfoque permite la detección,

localización y caracterización no supervisada de tumores, además de proporcionar datos de entrenamiento sintéticos para aumentar la robustez de los modelos para tareas posteriores, como la segmentación de tumores. En tercer lugar, se desarrolla un modelo de difusión latente multicondicional para transformar imágenes de MRI sin realce de contraste en imágenes sintéticas de DCE-MRI variables en el tiempo que localizan los tumores y predicen sus patrones cinéticos de realce de contraste. Abordando la necesidad de una métrica cuantitativa respectiva de evaluación, se propone la Distancia Radiómica de Fréchet (FRD) para medir la calidad de la imagen (sintética) basada en la variabilidad de biomarcadores. En cuarto lugar, se propone una red generativa adversarial condicionada a la malignidad de massas (MCGAN) para generar datos sintéticos como mecanismo de preservación de la privacidad para el entrenamiento de modelos de aprendizaje profundo sin datos de pacientes individuales. Mediante la comparación y la combinación con la privacidad diferencial, se demuestra que los datos sintéticos de mamografía mejoran el rendimiento en múltiples escenarios de clasificación de cáncer con preservación de la privacidad. En quinto lugar, se presenta la biblioteca medigan como plataforma compartida para modelos generativos preentrenados que permite a los investigadores generar datos sintéticos de alta calidad en diversas modalidades de imagen sin necesidad de acceder directamente a datos sensibles de pacientes. Además, los modelos generativos de medigan se analizan exhaustivamente basándose en la Fréchet inception distance (FID) utilizando tanto extractores de características específicos del dominio radiológico como extractores de características estándar invariantes del dominio.

En conclusión, esta tesis demuestra el potencial del aprendizaje profundo generativo para abordar los desafíos clave en las imágenes de cáncer, presentando métodos novedosos en la simulación de mejora del contraste, localización de tumores, clasificación de cáncer preservando la privacidad y herramientas para compartir y evaluar modelos generativos, abriendo el camino para la integración de estos modelos en la práctica clínica y el avance de la atención sanitaria tanto para los pacientes individuales como para la sociedad en general.

Palabras clave: Modelos Generativos Profundos, Análisis de Imágenes Médicas, Síntesis de Datos, Simulación de Mejora de Contraste, Aprendizaje Automático con Preservación de la Privacidad, Evaluación de Datos Sintéticos.

Contents

Declaration of Authorship	7
Acknowledgements	11
Abstract	15
Resum	17
Resumen	19
List of Figures	25
List of Tables	27
List of Publications	29
1 Introduction	35
1.1 Oncology and Cancer Imaging	35
1.2 Medical Physics and Imaging Techniques	37
1.3 Contrast Agents in Medical Imaging	38
1.4 Deep Learning	39
1.4.1 The Potential of Deep Learning in Cancer Imaging	39
1.4.2 Fundamentals, Architectures and Progress	41
1.4.3 Deep Generative Models	44
1.4.4 Generative Model Evaluation	45
1.4.5 Towards Deep Generative Models in Cancer Imaging	46
1.5 Research Background	47
1.5.1 Project Framework	47

1.5.2	Objectives	48
1.6	Contributions and Document Structure	49
2	Cancer Image Synthesis: Challenges, Solutions and Their Validation	53
2.1	Contribution: Medical Image Analysis 2023 (Review & Meta-Analysis	54
3	Synthesis of Multi-Sequence Dynamic Contrast-Enhanced Breast MRI using Generative Adversarial Networks	105
3.1	Contribution: JMI 2024 (DCE-MRI Synthesis)	106
4	Time-Conditioned Pre- To Post-Contrast Breast MRI Synthesis using Latent Diffusion Models	143
4.1	Contribution: MICCAI 2024 (CCNet & FRD)	144
5	Generative Models for Patient Privacy Preservation	159
5.1	Contribution: MICCAI DeepBreath Workshop 2024 (MCGAN)	159
6	Sharing Generative Models to Overcome Data Scarcity Issues in Medical Imaging	173
6.1	Contribution: JMI 2023 (medigan)	174
7	Results and Discussion	205
7.1	Contrast Localization and Characterization	205
7.1.1	Modeling of Contrast Patterns	205
7.1.2	Impact and Benefit of Contrast Prediction Applications	206
7.1.3	Extension of Modeling Capabilities and Input Data Sources . .	207
7.1.4	Measuring the Uncertainty of Synthetic Contrast	208
7.1.5	Detection of Anomalous Regions-of-Interest	209
7.1.6	Additional Longitudinal Temporal Pattern Prediction	210
7.2	Towards New Data and Model Sharing Paradigms	211
7.2.1	Solving Data Availability	211
7.2.2	Privacy-Aware Sharing of Models and Data	212
7.3	Trustworthy Artificial Intelligence in Healthcare: Synergizing FUTURE- AI and Generative Models	215
7.3.1	General: Privacy via De-Identifying Data Synthesis	215
7.3.2	Fairness: Via Balanced Dataset Generation	216
7.3.3	Robustness: Enhancement via Synthetic Data	217
7.3.4	Explainability: Via Controlled Variations in Input Conditions .	217

Contents

- 7.3.5 Explainability: Via Counterfactual Synthesis 218
- 7.3.6 General: Definition of Adequate Evaluation Plans 218
- 7.4 The Intricacies of Evaluating Generative Models 219
 - 7.4.1 Evaluating Synthesis Studies using SynTRUST 219
 - 7.4.2 Quality Assessment for Medical Image Synthesis 223
 - 7.4.3 Metric Contributions for Image Synthesis Assessment 225
 - 7.4.4 Emerging Generative Model Evaluation Paradigms 229
- 8 Conclusion 235**
- Bibliography 239**

List of Figures

1.1	Visualization of the structure of contents of the present thesis. An in-depth survey in Section 2 identifies current challenges and problems in cancer imaging and proposes approaches to address these problems based on generative modeling, while also introducing a novel meta-analysis framework to assess the trustworthiness of studies applying these approaches. The subsequent sections 3 - 6 present and evaluate various generative modeling methods and applications to define, analyze and address specific problems identified in Section 2.	52
7.1	Creativity in deep generative models: Illustrative visual depiction of interpolation and extrapolation capabilities of generative models. Out-of-distribution (OOD) samples containing concepts that are not readily present in the training data distribution, but are extrapolated abstractions thereof. The demonstrated images were generated by a DALL·E model using the ChatGPT 4o interface (version date: 12.11.2024). . .	231

List of Figures

List of Tables

7.1	Summary of the SynTRUST framework defined in Chapter 2 alongside its dimensions and building blocks for evaluating the trustworthiness of studies proposing synthetic data or respective data generation methods in medical imaging.	219
7.2	Summary, applications, and challenges of evaluation dimensions for synthetic data in medical imaging based on Chapter 3.	223
7.3	Overview of generative model evaluation metrics in medical imaging proposed in the present thesis across chapters 3, 4, and 6.	226

List of Publications

Authored and Accepted

The author contributed to the following peer-reviewed, subject-related first-authored publications (most recent first). The selected core publications of this dissertation are marked with a star ():*

1. (*) **Osuala, R.**, Lang, D.M., Riess, A., Kaissis, G., Szafranowska, Z., Skorpuko, G., Diaz, O., Schnabel, J.A., Lekadir, K. (2025). Enhancing the Utility of Privacy-Preserving Cancer Classification using Synthetic Data. In: *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care* (pp. 54-64). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77789-9_6
2. (*) **Osuala, R.**, Lang, D. M., Verma, P., Joshi, S., Tsirikoglou, A., Skorpuko, G., Kushibar, K., Garrucho, L., Pinaya, W.H.L., Diaz, O., Schnabel, J.A., Lekadir, K. (2024). Towards learning contrast kinetics with multi-condition latent diffusion models. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 713-723). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72086-4_67
3. **Osuala, R.**, Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W. H., Diaz, O., & Lekadir, K. (2024). Pre-to post-contrast breast MRI synthesis for enhanced tumour segmentation. In: *Medical Imaging 2024: Image Processing (Vol. 12926, pp. 226-237)*. SPIE. <https://doi.org/10.1117/12.3006961>
4. (*) **Osuala, R.**, Skorpuko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Diaz, O., Lekadir, K. (2023). medigan: a Python library of pretrained generative models for

List of Publications

- medical image synthesis. *Journal of Medical Imaging*, 10(6), 061403-061403. <https://doi.org/10.1117/1.JMI.10.6.061403>
5. (*) **Osuala, R.**, Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., Lekadir, K. (2023). Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Medical Image Analysis*, 84, 102704. <https://doi.org/10.1016/j.media.2022.102704>
6. Szafranowska, Z.[†], **Osuala, R.**[†], Breier, B., Kushibar, K., Lekadir, K., & Diaz, O. (2022). Sharing generative models instead of private data: a simulation study on mammography patch classification. In: *16th International Workshop on Breast Imaging (IWBI2022) (Vol. 12286, pp. 169-177)*. SPIE. <https://doi.org/10.1117/12.2625781> ([†] shared first authorship between SZ and OR)

Authored and Under Peer-Review

The author contributed to the following subject-related first-authored publications that are still under peer-review. The selected core publications of this dissertation are marked with a star ():*

7. (*) **Osuala, R.**, Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W.H.L., Lang, D.M., Schnabel, J.A., Diaz, O., Lekadir, K. (2024). Simulating Dynamic Tumor Contrast Enhancement in Breast MRI using Conditional Generative Adversarial Networks. Submitted to: *Journal of Medical Imaging* (Under Review). <https://doi.org/10.48550/arXiv.2409.18872>

Led and co-authored

The author also led the supervision of the following peer-reviewed, subject-related co-authored publications (most recent first):

8. Arcas, M. B., **Osuala, R.**, Lekadir, K., & Díaz, O. (2024). Mitigating annotation shift in cancer classification using single-image generative models. In *17th International Workshop on Breast Imaging (IWBI 2024) (Vol. 13174, pp. 512-523)*. SPIE. <https://doi.org/10.1117/12.3025548>

9. Kalb, T., Kushibar, K., Cintas, C., Lekadir, K., Diaz, O., & **Osuala, R.** (2023). Revisiting skin tone fairness in dermatological lesion classification. *In Workshop on Clinical Image-Based Procedures (pp. 246-255)*. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45249-9_24

Co-authored

The author also contributed to the following peer-reviewed co-authored publications (most recent first):

10. Dimitriadis, A., Kalliatakis, G., **Osuala, R.**, Kessler, D., Mazzetti, S., Regge, D., Diaz, O., Lekadir, K., Fotiadis, D., Tsiknakis, M. and Papanikolaou, N. (2025). Assessing Cancer Presence in Prostate MRI Using Multi-Encoder Cross-Attention Networks. *Journal of Imaging*, 11(4), p.98. <https://doi.org/10.3390/jimaging11040098>
11. Garrucho, L., Reidel, C.A., Kushibar, K., Joshi, S., **Osuala, R.**, Tsirikoglou, A., Bobowicz, M., del Riego, J., Catanese, A., Gwoździewicz, K., Cosaka, M.L., Abo-Elhoda, P.M., Tantawy S.W., Sakrana, S.S., Shawky-Abdelfatah, N.O., Salem, A.M.A., Kozana, A., Divjak, E., Ivanac, G., Nikiforaki, K. Klontzas, M.E., García-Dosdá, R., Gulsun-Akpınar, M., Lafcı, O., Mann, R., Martín-Isla, C., Prior, F., Marias, K., Starmans, M.P.A., Strand, F., Díaz, O., Igual, L., Lekadir, K. (2025). A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Scientific Data*. 12, 453. <https://doi.org/10.1038/s41597-025-04707-4>
12. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., Asselbergs, F.W., Prior, F., Collins, G.S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J.A., Kushibar, K., Riklund, K., Marias, K., Amugongo, L.M., Fromont, L.A., Maier-Hein, L., Cerdá Alberich, L., Martí-Bonmatí, L., Cardoso, M.J., Bobowicz, M., Shabani, M., Tsiknakis, M., Zuluaga, M.A., Fritzsche, M., Camacho, M., Lingurar, M.G., Wenzel, M., De Bruijne, M., Tolsgaard, M.G., Goisauf, M., Cano Abadía, M., Papanikolaou, M., Lazrak, N., Pujol, O., **Osuala, R.**, Napel, S., Joshi, S., Klein, S., Aussó, S., Rogers, W.A., Puig-Bosch, X., Salahuddin, Z., Starmans, M.P.A., and the FUTURE-AI Consortium (2025). FUTURE-AI: International consensus guide-

List of Publications

- line for trustworthy and deployable artificial intelligence in healthcare. *British Medical Journal (BMJ)*. 388. <https://doi.org/10.1136/bmj-2024-081554>
13. Garrucho, L., Delegue, E., **Osuala, R.**, Kessler, D., Kushibar, K., Diaz, O., Lekadir, K., Igual, L. (2025). Fat-Suppressed Breast MRI Synthesis for Domain Adaptation in Tumour Segmentation. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care* (pp. 202-211). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77789-9_20
 14. Daum, D., **Osuala, R.**, Riess, A., Kaissis, G., Schnabel, J. A., Di Folco, M. (2024). On Differentially Private 3D Medical Image Synthesis with Controllable Latent Diffusion Models. In: *MICCAI Workshop on Deep Generative Models* (pp. 139-149). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72744-3_14
 15. Fischer, S. M., Felsner, L., **Osuala, R.**, Kiechle, J., Lang, D. M., Peeken, J. C., Schnabel, J. A. (2024). Progressive Growing of Patch Size: Resource-Efficient Curriculum Learning for Dense Prediction Tasks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 510-520). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72114-4_49
 16. García, E., Lladó, X., Mann, R. M., **Osuala, R.**, Martí, R. (2024). Breast composition measurements from full-field digital mammograms using generative adversarial networks. In: *17th International Workshop on Breast Imaging (IWBI 2024)* (Vol. 13174, pp. 181-188). SPIE. <https://doi.org/10.1117/12.3026925>
 17. Joshi, S., **Osuala, R.**, Garrucho, L., Tsirikoglou, A., del Riego, J., Gwoździewicz, K., Kushibar K., Diaz O., Lekadir, K. (2024). Leveraging epistemic uncertainty to improve tumour segmentation in breast MRI: an exploratory analysis. In: *Medical Imaging 2024: Image Processing* (Vol. 12926, pp. 292-300). SPIE. <https://doi.org/10.1117/12.3006783>
 18. Garrucho, L., Kushibar, K., **Osuala, R.**, Diaz, O., Catanese, A., Del Riego, J., Bobowicz M., Strand F., Igual L., Lekadir, K. (2023). High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12, 1044496. <https://doi.org/10.3389/fonc.2022.1044496>

19. Joshi, S., **Osuala, R.**, Martín-Isla, C., Campello, V. M., Sendra-Balcells, C., Lekadir, K., & Escalera, S. (2021). nn-UNet training on CycleGAN-translated images for cross-modal domain adaptation in biomedical imaging. In: *International MICCAI Brainlesion Workshop (pp. 540-551)*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-09002-8_47
20. Diaz, O., Kushibar, K., **Osuala, R.**, Linardos, A., Garrucho, L., Igual, L., Radeva P., Prior F., Gkontra P., Lekadir, K. (2021). Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica medica*, 83, 25-37. <https://doi.org/10.1016/j.ejmp.2021.02.007>

List of Publications

Chapter 1

Introduction

In this chapter, the background and context is outlined in which the present doctoral thesis is embedded, emphasizing the significance of the topics of the thesis within its broader context. Furthermore, the established objectives driving the research contributions are described alongside the projects that have benefited from the findings of this thesis. Lastly, a concise overview of the structure of the subsequent chapters is provided to offer the reader a clear roadmap throughout the thesis.

1.1 Oncology and Cancer Imaging

Cancer, as one of the leading causes of mortality worldwide, is a complex, heterogeneous disease characterized by uncontrolled cell growth and the potential to invade other parts of the body, resulting in substantial morbidity and mortality. Oncology as a specialized field of medicine is dedicated to the study, diagnosis, treatment, and prevention of cancer encompassing the understanding of molecular and cellular cancer mechanisms, as well as the definition of screening programs, interventions, and personalized treatment strategies, for instance, based on genetic and clinical profiling. To date, cancer continues to rank among the leading causes of morbidity and mortality worldwide [67], with an estimated global incidence of 19.98 million and a mortality of 9.74 million in 2022 [226]. In 2018, lung cancer accounted for the highest number of deaths (1.76 million), followed by colorectal (0.86 million), stomach (0.78 million), liver (0.78 million), and breast cancer (0.63 million) [225]. In 2022, with the exception of stomach and liver cancer, these numbers have increased notably with lung cancer accounting for a mortality of 1.82 million, followed by colorectal (0.90 million), liver (0.76 million), breast (0.67 million),

1.1. Oncology and Cancer Imaging

and stomach cancer (0.66 million) [226]. In the future, it is expected that the quantity of cancer cases further increases due to ageing and world population growth [101]. The aforementioned global burden posed by cancer on patients and healthcare systems could be reduced by earlier detection as well as more effective treatments [101]. As each tumor undergoes several stages of critical genetic alterations to become cancerous and thereby often becomes less treatable, earlier detection bears potential to intervene before such alterations have occurred. Based on these genetic alterations of cancerous cells they obtain growth advantages compared to regular cells in a process related to Darwinian evolution [87]. With the aim of providing an "organizing principle for rationalizing the complexities of neoplastic disease", Hanahan et al [87, 88] specify the hallmarks of genetic alterations of cancer cells. These include resisting programmed cell death (apoptosis), sustaining proliferative signaling, evading growth suppressors, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis, as well as the emerging hallmarks of energy metabolism reprogramming, and immune destruction evasion. Both inflammation and genome instability, with the latter promoting genetic diversity, are key drivers influencing the rate at which these hallmarks are acquired. As a further dimension of complexity, seemingly regular non-cancerous cells are recruited within the tumor microenvironment (TME) that then contribute to obtaining hallmark traits. With multiple tumor sub-populations within the TME it becomes inherently complex and difficult to predict which tumor cell subpopulations outcompete others to become the dominant lineage in the evolutionary trajectory. As a result, tumor lesions exhibit continuous phenotypic and genotypic evolution [20], with the potential to accumulate additional deleterious mutations at any time, further exacerbating the complexities of treatment and prognosis. From a phenotypic perspective, solid tumors, after evolving from a single mutated cell, typically follow a Gompertzian growth pattern [164, 71]. They then often become detectable through medical imaging modalities only after reaching an approximate size of 10^9 cells ($\approx 1\text{cm}^3$). For human cancer cell lines, in vitro studies have suggested a theoretical limit for detection of approximately 10^5 to 10^6 cells using PET imaging. However, in clinical settings, this threshold is significantly higher, influenced by factors such as the specific cancer type, the background noise and radiation, as well as the characteristics of the cell line at hand [69, 71]. The vast variability across cancer types and patients in terms of tumor morphology, location, and progression underscores the importance of sophisticated diagnostic tools able to capture subtle differences in tissue composition and structure, for instance to facilitate the detection, characterization, and differentiation of malignant and benign lesions. Currently, medical imaging remains the primary

diagnostic tool for detecting and diagnosing tumors, providing non-invasive methods to visualize abnormalities within the body. In particular, radiologic imaging modalities are indispensable in clinical oncology, where radiologists rely on visual assessment of magnetic resonance imaging (MRI), computed tomography (CT), ultrasound (US), x-ray modalities, and positron emission tomography (PET) [71, 100, 146]. Screening mammography (MMG), which is based on x-ray techniques, is essential for the early detection of breast cancer and, thus, linked to a reduction in breast cancer mortality [3].

1.2 Medical Physics and Imaging Techniques

In contemporary clinical practice, the ability to visualize internal body structures through medical imaging has become a standard procedure, offering critical diagnostic and prognostic insights that are fundamental to modern medicine. Despite earlier work by Goodspeed and Jennings [211], it was, however, not until the faithful day of November 8, 1895, on which Wilhelm Conrad Roentgen discovered what he called *X-rays* during experiments with a Crookes cathode ray tube [19]. While radiographs were the sole form of medical imaging for decades, this progress lead the way for the discovery of additional imaging techniques with different qualities and based on different medical physics technologies. After Alessandro Vallebona invented conventional tomography moving the X-ray source and X-ray detector in tandem to focus on an object of interest, CT was developed with significant contributions from Sir Godfrey Hounsfield and Alan M. Cormack by acquiring radiographic projections from multiple angles for reconstruction of a two-dimensional image using mathematical models. Often co-registered with CT for enhanced anatomical context, further nuclear medicine modalities, such as Single-photon emission computed tomography (SPECT) and PET, use radioactive tracers to detect gamma rays in SPECT or to map metabolic activity by registering gamma rays emitted from PET. A further radiation-free, cost-effective, and widely-accessible alternative to CT is ultrasound, which transmits high-frequency sound waves into the human body to then create images by mapping the echoes based on their amplitude and time delay. While CT images facilitating unprecedented diagnostic accuracy, the innovation of MRI, with first images published in 1973 [123], further advanced the field of medical imaging and oncology towards improved capturing of soft tissue contrast and anatomic detail. MRI is based on a magnetic field around the patient in which radiofrequency pulses excite protons in the patient's body before the emitted signals during relaxation of these protons are captured and reconstructed into

1.3. Contrast Agents in Medical Imaging

an image [19, 211, 63]. The complementary use of various imaging modalities in cancer diagnostics leverages the unique strengths of each individual technique. For example, mammography continues to be the gold standard for breast cancer screening, offering high sensitivity and population-wide scale for early detection screening programs. In contrast, MRI's superior soft-tissue contrast makes it indispensable for identifying cancers in the brain, liver, and prostate. PET imaging, with its ability to provide detailed metabolic information, is frequently combined with CT or MRI to achieve comprehensive cancer characterization. Such integrated approaches, accompanied by further biopsies and clinical tests, enable multidimensional understanding of cancer, supporting more accurate diagnoses, precise staging, and effective treatment planning.

1.3 Contrast Agents in Medical Imaging

With these advances in imaging techniques, contrast agents (CA) became increasingly popular with the first designed for MRI becoming available for clinical use in 1988 [141]. CAs are typically administered intravenously and comprise substances such as superparamagnetic iron oxide particles and soluble paramagnetic metal chelates (e.g., gadolinium), which enhance the visibility of specific structures in imaging, such as tumors, by exploiting variations in contrast-enhanced blood flow and vascular permeability [232, 141]. Types of CAs can be categorized by their effect on the image, magnetic behavior and biodistribution in the body. CAs have been revolutionizing the precision and utility of contrast-enhanced MRI (CE-MRI) healthcare demonstrated by over 300 million procedures performed globally for improved diagnostics and therapy monitoring [232, 141]. Currently, contrast media are used in approximately 25% of MRI examinations, particularly for imaging the brain, spine, abdomen, breast, heart, and for MR angiography (MRA) [141]. Despite their successes, CA administration is associated with multiple risks and drawbacks including allergic and physiologic reactions, kidney complications (especially in patients with renal impairment), nephrogenic systemic fibrosis, and the deposition of CA substances in the body with unknown long-term health consequences [144, 165, 98, 108, 163, 166], apart from Symptoms Associated with Gadolinium Exposure (SAGE) with unknown causal relationship to Gadolinium-based CA [148]. In fact, in response to a 2016 request from the European Commission, the European Medicines Agency (EMA) recommended restricting certain intravenous linear gadolinium-based contrast agents (GBCAs) to reduce the potential risks of gadolinium deposition [65]. In addition, CA administration, which is cumbersome and unpleasant for patients, increases the complexity and also the cost of imaging

procedures. The usage of CA has also been shown to pose environmental concerns due to the corresponding pollution of aquatic ecosystems including drinking water supplies [180, 48, 166], with additional health risks stemming from its respective degradation products [29]. To this end, the advent of deep learning offers potential in optimizing contrast-enhanced imaging by predicting and simulating contrast uptake patterns, either by reducing dosages of GBCAs or by replacing them with synthetic contrast-enhanced imaging while maintaining diagnostic accuracy. By generating synthetic contrast-enhanced images from non-contrast data, deep learning methods can minimize patient exposure to CA and their associated risks, in particular in cases where CA injections are contraindicated [166, 158, 157, 236, 191].

1.4 Deep Learning

1.4.1 The Potential of Deep Learning in Cancer Imaging

In general, despite the strengths of traditional imaging methods, cancer diagnosis through imaging faces several critical limitations such as intra- and inter-observer variability, cancer heterogeneity and intra- and inter-tumor variability limiting standardized diagnostic criteria, as well as high diagnostic error rates, and variations in imaging protocols, imaging artifacts, diagnostic uncertainty, reproducibility, consistency [162, 20, 93, 97, 62, 223, 28, 100]. Furthermore, the acquisition and evaluation of medical imaging data is a time-intensive and resource-demanding process, making it inherently costly and requires assessment by trained professionals with specialist knowledge and experience. Moreover, the previously described progresses in medical imaging technologies have significantly increased the volume of acquired imaging data and, thus, the amount of data requiring professional analysis [147], placing a burden on the workload of radiology departments and further exacerbating the variability of diagnostic quality across radiology assessments stemming from a radiologist’s level of expertise, experience, and the time dedicated interpretation of a given image [223, 100, 62]. Given these constraints, the appeal of automated and semi-automated (e.g., triaging, second or third opinion) becomes apparent due to its potential of increasing diagnostic accuracy, reducing costs, streamlining workloads, and assist radiologists (e.g., radiologists in training) in clinical decision-making processes. Such computer-assisted tools for detection, diagnosis, and treatment planning and monitoring encompass traditional machine learning approaches as well as cutting-edge deep learning methods.

What is commonly referred to as *traditional machine learning* in medical imaging,

1.4. Deep Learning

are approaches that rely on handcrafted feature extraction, requiring domain expertise to quantify image characteristics in contrast to deep learning where hierarchical features are learned directly from raw data. One example of such handcrafted features are Radiomics features. Introduced in 2012 [121, 118], Radiomics have been predominantly designed for and applied to radiological imaging data in oncology but has been shown to be also useful in predictive tasks in other medical imaging domains such as cardiology [145]. It involves extracting quantitative features by converting digital medical images into structured data. Such data describes, among others, the shape, texture, intensity, and corresponding statistical patterns of regions-of-interest within the image [213] that are not only often imperceptible to the human eye, but also quantify imaging biomarker relevant for disease detection and characterization. Traditionally, these handcrafted radiomics features serve as inputs for machine learning models such as linear regression models, random forests, or support vector machines. These models then classify or predict outcomes by identifying patterns in the extracted features. Since these predefined features depend heavily on domain expertise, it is a significant assumption that they encompass all relevant patterns and nuances necessary to capture the complexities of imaging and disease characteristics across various modalities and conditions to accurately predict the desired outcomes. The shift from handcrafted features such as radiomics towards deep learning, where feature extraction is part of the learning algorithm, holds immense additional potential due to its capability of learning and interpreting features and patterns of arbitrary complexity from the training data. This not only enables more nuanced prediction capabilities, but can also increase the consistency and generalizability across datasets with high sample diversity. Improved generalizability is particularly useful within personalized oncology considering the heterogeneity of cancer genotypes and phenotypes and the corresponding multi-source, multi-modal and multi-dimensional data settings.

In recent years, advancements in artificial intelligence (AI) and deep learning have shown promise in addressing challenging medical imaging problems, particularly through the use of computer-aided detection and diagnostic (CAD) systems [20, 53, 129, 3, 93, 79]. Deep learning models can support radiologists by improving diagnostic accuracy, reducing variability, and enhancing decision-making [49, 20]. As an example, recent studies on screening mammography have demonstrated that complementarily integrating deep learning models with radiologist assessments achieves the highest performance in cancer detection [188, 186, 149]. By way of illustration, Salim et al [186] showed that an externally validated breast cancer detection model can achieve more accurate predictions compared to first-reader radiologists in evaluating screening

mammograms; nonetheless, the highest positive breast cancer detection rate was accomplished by combining the deep learning model’s predictions with the first-reader radiologist assessments. In general, there is vast potential for deep learning to decrease error rates in diagnostic image interpretation, as exemplified by Newman et al [162] surveying diagnostic errors and reporting a false negative rates for cancer of 11.1%. To this end, any significant advancements in deep learning are likely to directly enhance computer-aided detection and diagnosis (CAD) systems, enabling more precise workflows for screening and treating cancer-related diseases and ultimately improving clinical outcomes.

1.4.2 Fundamentals, Architectures and Progress

Deep learning relies on artificial neural networks composed of multiple layers, enabling the extraction of hierarchical and complex features from input data, such as images. These features are mapped into a lower-dimensional latent space that can learn useful multi-faceted decision boundaries for prediction tasks, including classification and segmentation. This latent feature space is iteratively refined during model training through optimization techniques, such as backward propagation of errors (backpropagation) and gradient descent, to improve model accuracy and performance. Gradient descent is the foundation of most deep learning optimization algorithms, as it adjusts a deep learning model’s trainable parameters by using the gradient of the loss function to update these parameters in the direction that minimizes the loss. Backpropagation computes these gradients by applying the chain rule of calculus determining how much each parameter contributes to the error defined by the loss function. The model’s parameters are the trainable weights and biases of each node of each layer of the artificial neural network. In general, these weights serve as multipliers in sequential matrix multiplications, where the outputs from the previous nodes (or the neural network’s input data) are multiplied by the current node’s weights, optionally followed by the addition of a bias term. Typically, the resulting value is passed through a non-linear activation function (e.g., a sigmoid function), which is essential for enabling the network to model and learn complex patterns.

For high-dimensional data, such as images used as input for neural networks, a specialized variation of linear neural networks—known as the convolutional neural network (CNN)—has demonstrated exceptional performance. Initially introduced as the Neocognitron by Kunihiko Fukushima in the 1980s [72, 73], this architecture was designed as a vision-focused neural network with hierarchical layers. It was further

1.4. Deep Learning

significantly advanced by Yann LeCun et al. [126] in 1989, who proposed the use of backpropagation as an efficient method for training vision-based convolutional neural networks in a supervised learning framework, while also suggesting [125] to carefully control the number of trainable vision neural network to enhance their generalization capability despite high input dimensionality.

Since then, further advancements in both methodology and hardware, such as graphics processing unit (GPU) technology, have substantially increased the ability to train deep learning algorithms, including CNNs [127] on large-scale image datasets enabling groundbreaking results in computer vision [41, 115] and cancer imaging [42]. Notably, the success of AlexNet during the 2012 ImageNet challenge [115] served as a turning point, spurring the widespread application of deep neural networks to a variety of challenges across multiple domains, including medical imaging [193, 53, 138]. For instance, the U-Net architecture [183] is a widely used convolutional neural network in computer vision and medical imaging, particularly for image segmentation and image-to-image translation tasks. It features a symmetric encoder-decoder structure, where the encoding path extracts hierarchical features, and the decoding path restores spatial resolution, with skip connections between encoder and decoder layers preserving fine-grained details. Building on this framework, nnU-Net [99] automates architectural configuration, preprocessing, and training schedules based on dataset characteristics, eliminating the need for manual tuning and thereby achieving state-of-the-art performance in multiple challenging medical image segmentation tasks.

In 2017, the transformer architecture was introduced by Vaswani et al.[215], initially revolutionizing the field of natural language processing [51] and subsequently extending its impact to computer vision[57]. Unlike convolutional neural networks (CNNs), the transformer architecture relies on a self-attention mechanism, enabling it to model long-range dependencies and complex relationships within the data. This approach facilitates fast parallel processing, as the architecture does not depend on sequential operations typically required in recurrent networks or the fixed-size receptive fields of CNNs. For vision tasks, such as in the Vision Transformer (ViT) [57], input images are divided into fixed-size patches, which are flattened and embedded into a sequence of tokens via a linear projection. Positional embeddings are added to retain spatial information. The transformer then computes self-attention across all token pairs, capturing intricate patterns and global dependencies between the image patches. This design allows transformers to efficiently learn hierarchical and context-aware representations, enabling state-of-the-art performance across a wide range of computer vision tasks. Unlike convolutional neural networks (CNNs), transformers do not contain

an inductive locality bias, meaning their architecture does not assume that related features are spatially close within an image. Subsequent hierarchical vision transformer architectures, such as the Swin Transformer [140], combine some of the benefits of both architectures by introducing a shifted window mechanism, which restricts self-attention computation to non-overlapping local windows while enabling cross-window connections for capturing global context.

Recently, foundational models have emerged as a transformative paradigm in deep learning, where large-scale pretrained architectures learn generalizable representations that enable effective transfer across diverse tasks and domains [24, 10]. These models are typically pretrained on extensive, multimodal datasets and fine-tuned for specific applications, reducing the reliance on task-specific annotated data, which is often scarce in medical imaging [11, 154]. Leveraging advances in vision and language processing, foundational models popularize methodological innovations such as textually-prompted models, visually-prompted models, and multimodal frameworks, facilitating tasks ranging from segmentation and classification to anomaly detection and radiology report generation. In the context of medical imaging, these models are particularly promising due to their ability to integrate and adapt to multimodal data, such as combining radiological images with clinical reports or genomic data [154, 11, 197]. By training on expansive datasets, foundational models like Segment Anything Model (SAM) [112] have demonstrated advanced capabilities in generalizing across medical imaging tasks including the segmentation of complex multimodal anatomical structures.

Despite its growing importance in cancer imaging workflows, deep learning applications faces significant challenges in the domain of radiology [20]. These challenges include limited access to publicly available datasets due to privacy regulations and institutional constraints [147], as well as a high cost and labor-intensive process of annotating medical data, hindering the development of robust and generalizable models. Additionally, domain shifts caused by variations in imaging protocols, equipment, and patient demographics reduce model performance across different clinical settings. These challenges allude to the potential of generative models to address data scarcity, data privacy concerns, and model robustness issues [206, 166, 6, 79]. For instance, generated synthetic data can lead to more robust and privacy-preserving model training while generative models can also translate images from one domain to another to mitigate domain shifts and differences between clinical settings.

1.4. Deep Learning

1.4.3 Deep Generative Models

In machine learning, generative and discriminative models serve distinct purposes and diver in how they process and model data. Discriminative models learn a conditional probability $p(y|x)$, where y is the target label and x is the input data sample. Often implemented as convolutional or transformer neural networks, they learn a high-dimensional decision boundary in latent space z that separates data points x with different labels y from each other. On the other hand, generative models learn the underlying probability distribution of the training data $p(x)$. This enables them to then generate synthetic data samples that resemble the original data by sampling a point x_{gen} from the learned data distribution $p(x)$. The probability distribution that generative models learn can further also be a conditional distribution $p(x|y)$, where x represents the data samples, and y provides an additional context or constraint for the generation process. While generative models are particularly useful for creating synthetic data, learning latent representations, and handling unsupervised tasks, discriminative models are well suited for supervised tasks such as classification, segmentation, or regression, that require mapping inputs to specific outputs. When used complementarily, generative models can augment training datasets for discriminative models, often resulting in an enhancement of performance and/or robustness, which is particularly relevant in the domain of medical imaging due to its scarcity of large datasets x and/or respective annotations y .

Several modeling methodologies and neural network architectures have been proposed to train deep generative models, each bearing their own unique strengths: By way of example, Variational Autoencoders (VAEs) [111] consist of an encoder network that maps input data to a low-dimensional latent distribution (typically Gaussian) and a decoder network that then reconstructs the data sample from the encoded latent samples using a combination of a reconstruction loss and a regularization term (Kullback-Leibler divergence). In Generative Adversarial Networks (GANs) [83], two neural networks, a generator and a discriminator, are trained simultaneously in a competitive framework using a minimax loss function. The generator learns to transform random noise into samples from the data distribution, while the discriminator attempts to distinguish between real and generated samples. Inspired by thermodynamic diffusion processes, Denoising Probabilistic Diffusion Models (DDPMs) [199, 201, 92] iteratively reverse a predefined noising process by progressively denoising the data sample through a series of timesteps, thereby systematically recovering data from noise. The latent diffusion model (LDM) [181] counterparts of DDPMs directly denoise the latent representation

of encoded input data samples rather than the samples themselves. DDPMs and LDMs typically consists of a neural network, such as a U-Net [183], which predicts the noise added to the data at particular timesteps and is trained by minimizing the difference between the true and predicted noise using, e.g., using a mean squared error loss term. Furthermore, Normalizing Flows [179, 55, 56] allow for precise likelihood estimation and efficient data sampling by using invertible transformations to explicitly model complex data distributions. Autoregressive models, such as large language models (LLMs), vision-language models (VLMs), and vision transformers [57, 215], generate data by modeling each data point, e.g. a subset of a sample, as a conditional probability in dependence of the preceding or surrounding points. This stepwise synthesis process implicitly allows to learn complex probabilistic dependency structure within the data, allowing the model to generate outputs that preserve both local and global coherence.

1.4.4 Generative Model Evaluation

The evaluation of generative models is often based on the synthetic data generated by these models. One key indicator is the utility of the synthetic data to serve specific purposes such as improving downstream task results of a discriminative model, when that discriminative model is trained on (additional) synthetic training data. Apart from downstream task based evaluation, the computer vision domain has introduced several methods for evaluating synthetic data, often using real data as a means for comparison. Such methods, which use image-to-image comparisons include the mean squared error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), multi-scale (MS) structural similarity index measure (SSIM) [219], as well as the Learned Perceptual Image Patch Similarity (LPIPS) [235]. Other methods such as the Fréchet Inception Distance (FID) [89] or the Kernel Inception Distance (KID) [21] compare the distribution of the generated data with the distribution of the real data. Recent developments in generative model evaluation highlight the importance of addressing the numerous limitations of these traditional measures [26, 27]. In particular when evaluating synthetic medical images, conventional metrics based on natural images such as FID or LPIPS are not specifically designed for the characteristics and modalities of the medical imaging domain. By way of illustration, distinct lesion sizes, small artifacts or subtle, clinically relevant deviations in synthetic medical images can influence clinical interpretation as well as the features learned from such images in downstream task model training. Metrics for medical image evaluation are to further incorporate clinical relevance, which includes accurate reflection of the commonly highly

1.4. Deep Learning

heterogeneous features of anatomy and disease. These features are often only known in part and only by experienced domain experts. However, manual inspection by trained clinicians, such as radiologists or pathologists, is subjective and resource-intensive. This underscores the need for domain-specific automated evaluation tools assessing synthetic medical data across its multiple dimensions including fidelity, diversity, adherence to provided input conditions, clinical utility, patient privacy, as well as biases and fairness, as is discussed in detail in Chapters 2 and 3.

1.4.5 Towards Deep Generative Models in Cancer Imaging

As previously stated, the application of deep learning in cancer imaging is constrained by the need for large, annotated datasets, which are often difficult to obtain due to privacy concerns, data heterogeneity, and the high costs associated with expert labeling. Generative models hold particular promise due to their potential of addressing these key challenges including data scarcity, privacy concerns, and the need for robust models capable of generalizing across diverse clinical settings. The generation of high-quality synthetic data converging towards the statistical properties of real imaging datasets poses a solution to the scarcity of annotated oncological datasets. By way of example, generative cancer imaging models can be applied to augment training datasets to improve downstream task model performance, realistic tumor appearances and variations can be synthesized, and privacy-preserving model development and training can be enabled. A further application is to use the distribution learned by generative models to scrutinize or edit imaging data revealing clinically relevant characteristics in an unsupervised manner. Examples of the latter applications include synthetic tumor contrast enhancement, where a generative model’s translation from non-contrast enhanced to contrast-enhanced images can highlight and thereby localize a tumor without the need of providing annotated tumor areas to the generative model [166, 191, 158, 157]. Related to that, a second example is anomaly detection, where, in a process often referred to as *normative learning* [38, 18], a generative model learns the distribution of healthy (i.e. normal) patient samples, which then allows to automatically detect non-healthy (i.e. abnormal) samples during inference. This detection is often based on comparing how the generative model, such as autoencoder [169], transformer [171], GAN [189], or diffusion models [170, 17, 229, 222, 85], adjusts the provided abnormal out-of-distribution sample to its normal within-distribution counterpart. Furthermore, multimodal generative models can integrate imaging with other patient data, such as genomic, textual, or clinical information, to create richer, context-aware

diagnostic tools that take into account the high variation of different modalities and data types in medicine.

Despite their promise, deep learning including deep generative models in cancer imaging face several challenges. These challenges are to be analyzed and documented in-depth to allow for a principled proposal of suitable and sustainable solutions. Such challenges include ensuring the clinical validity and applicability of synthetic data going beyond standard technical evaluation metrics from the domains of computer vision based on natural images [151, 224, 37]. Furthermore, the potential of the multiple uses of deep generative models and synthetic cancer imaging data is to be assessed and validated in a reproducible manner to detect beneficial clinical applications and uncover quantifiable knowledge about the benefit of respective generative modeling approaches.

1.5 Research Background

1.5.1 Project Framework

The research presented in this thesis was conducted at the Barcelona Artificial Intelligence in Medicine Laboratory (BCN-AIM) at the Department of Mathematics and Computer Science at the University of Barcelona (UB) from 2020 until 2025. Part of the research of this thesis was conducted during a research stay from 10/2023 until 08/2024 at the laboratory of Professor Julia A. Schnabel, namely, the Institute for Computational Imaging and AI in Medicine (CompAI) at Technical University Munich and at the Institute of Machine Learning in Biomedical Imaging (IML) at Helmholtz Center Munich.

- From 10/2023 until 03/2023, the research stay was funded by the Montcelimar doctoral student research stay grant co-financed by Fundació Montcelimar and University of Barcelona.
- From 05/2024 until 07/2024, this research stay was funded by a research stay grant from the Helmholtz Information and Data Science Academy (HIDA).

Furthermore, the research described in this PhD thesis was carried out within the frameworks of the following projects.

1. [2020 - 2025] EuCanImage: “A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and

1.5. Research Background

Precision Medicine in Oncology”. Awarded in 2020 under the Horizon 2020 research and innovation programme by the European Commission (EC). Signature Date: 29 May 2020. Ref: 952103

2. [2022 - 2025] RadioVal: “International Clinical Validation of Radiomics Artificial Intelligence for Breast Cancer Treatment Planning”. Awarded in 2022 under the Horizon Europe research and innovation programme by the European Commission (EC). Signature Date: 19 May 2022. Ref: 101057699
3. [2022 - 2025] FUTURE-ES: “Trustworthy and multi-centre artificial intelligence for future medical imaging in Spain”. Awarded in 2023 by the Ministry of Science, Innovation and Universities of Spain. Ref: PID2021-126724OB-I00
4. [2024 - 2025] AIMED: “Reliable Artificial Intelligence to improve medical imaging workflows”. Awarded in 2024 by the Ministry of Science, Innovation and Universities of Spain. Ref: PID2023-146786OB-I00

1.5.2 Objectives

The overall objective of this thesis is to develop, validate, and apply medical image synthesis to address critical challenges in cancer imaging. As part of the EuCanImage, RadioVal, FUTURE-ES and AIMED projects, the goal of the present work encompasses the application, experimentation and validation of trustworthy artificial intelligence in radiology focusing on the realm of breast cancer and can be summarized as follows.

To develop machine learning frameworks for the synthesis of breast cancer imaging modalities, such as mammography and MRI, and to assess automated image analysis methods for tumor diagnosis and interpretation in radiology, contributing towards a European cancer imaging platform for enhanced artificial intelligence in oncology.

To define and meet the multiple requirements for developing automated machine learning frameworks and methods in breast cancer imaging, the present work is structured into a series of sub-objectives. These sub-objectives are designed to ensure a step-by-step approach to achieving the overarching goal of advancing generative and deep learning methods for cancer image analysis:

- **The symbiosis of generative modeling and cancer imaging challenges:**
To identify the current state of issues, challenges and solutions in cancer imaging

and generative models, and to concisely structure them into lines of research that provide a roadmap for the use of generative models in cancer imaging.

- **Meta-analysis of trustworthy generative models:** To design a principled medical imaging meta-analysis framework to assess the reliability, AI trustworthiness, and maturity of generative modeling studies and solutions.
- **Domain translation to improve tumor analysis in breast MRI:** To develop generative models for translating breast MRI between medical imaging modalities, such as converting non-contrast-enhanced to contrast-enhanced image sequences, to improve the characterization of cancerous lesions.
- **Privacy-preserving generative deep learning:** Investigate the utility of synthetic data generated by deep generative models as a tool for preserving patient privacy in medical imaging and compare and combine it with proven and established privacy mechanism such as differential privacy [60].
- **A multi-modal image synthesis toolbox:** To design and implement a toolbox to generate realistic and diverse cancer imaging datasets across modalities such as MRI and mammography, leveraging generative deep learning architectures and addressing data scarcity, privacy concerns, and domain variability.
- **Conduct and innovate synthetic data evaluation:** To conduct experimental validation of the utility of synthetic datasets. Overcome limitations of existing generative model evaluation metrics and propose and validate novel domain-specific methods for evaluating synthetic medical images.

1.6 Contributions and Document Structure

Accurate detection, localization, and characterization of cancerous lesions are essential for early diagnosis and effective treatment. Yet, this remains challenging due to, e.g., tumor heterogeneity and inter-observer variability, which complicate the distinction between healthy and cancerous tissue. While deep learning architectures such as convolutional and transformer-based networks show promise by capturing complex imaging patterns such as cancerous features, their clinical adoption is hindered by multiple challenges, among others, data scarcity, manual labeling requirements, sensitivity to domain shifts and imaging variability.

This thesis identifies and addresses the challenges in deep learning for cancer imaging through five core publications that propose novel frameworks, applications, and methods.

1.6. Contributions and Document Structure

These publications explore and ideate solutions based on deep generative models in cancer imaging, addressing the aforementioned challenges including, but not limited to, mitigating data and annotation scarcity, preserving patient privacy, overcoming data sharing restrictions, and characterization and localization of heterogeneous tumor manifestations. To this end, the present work alludes to the potential and versatility of generative models as a cornerstone of computer-aided methods capable of improving outcomes for individual patients as well as for society at large. Figure 1.1 outlines the structure of this thesis, highlighting the identified and categorized challenges in cancer imaging alongside their corresponding methodological solutions and applications. The thesis also includes an in-depth discussion of the presented contributions and the future research directions they inspire. The key contributions of this work are summarized as follows.

- **Chapter 2. Cancer Image Synthesis: Challenges, Solutions and Their Validation.** A large-scale survey of generative models in cancer imaging is conducted leading to the identification of key challenges and problems in the field. Furthermore, the literature is surveyed to find potential solutions, which are then concisely tailored to the identified challenges. Lastly, based upon the experiences and conclusions drawn from these surveys, the meta-analysis framework *SynTRUST* is derived, defined and applied to assess the trustworthiness and maturity level of cancer image synthesis studies and solutions.
- **Chapter 3. Synthesis of Multi-Sequence Dynamic Contrast-Enhanced Breast MRI using Generative Adversarial Networks.** Conditional generative adversarial networks (GANs) are applied to the challenges of cancer image scarcity and limited tumor annotations, by generating dynamic contrast-enhanced breast magnetic resonance imaging (DCE-MRI) sequences from their pre-contrast counterparts. With tumors commonly presenting as hypo- or hyper-intense in the DCE-MRI sequences, this generative modeling application enables unsupervised tumor detection, localization, and characterization, along with providing synthetic training data for enhancing tumor segmentation downstream task models. Experiments not only demonstrate increased robustness of tumor segmentation models trained with synthetic data, but also a novel investigation into modeling the temporal patterns by synthesizing multiple, in practice sequentially acquired, DCE-MRI sequences.
- **Chapter 4. Time-Conditioned Pre- To Post-Contrast Breast MRI Synthesis using Latent Diffusion Models.** A novel multi-conditional latent

diffusion models is developed to translate non-contrast enhanced MRI images into variable time-dependent synthetic DCE-MRI images localizing tumors and predicting their contrast enhancement kinetic patterns. Further, the need for respective quantitative image metrics is addressed by introducing the novel Fréchet Radiomics Distance (FRD) measuring synthetic image quality based on biomarker variability. FRD effectively addresses the limitations of the synthetic data evaluation methods from the computer vision domain, such as the FID, that are commonly applied in medical imaging. FRD not only provides interpretable features capable of explaining and pinpointing the metric's output values to particular imaging characteristics and biomarkers, but is also enables region-of-interest conditioned 2D as well as 3D medical imaging specific synthetic data evaluation.

- **Chapter 5. Generative Models for Patient Privacy Preservation.** A malignancy-conditioned generative adversarial network (MCGAN) is proposed to generated synthetic data conditioned on generating either malignant or benign mammography tumor masses. The synthetic data is further used to train a mass malignancy classification model and validated as privacy-preservation mechanism for training deep learning models without using data attributable to particular individual patients. Via comparison and combination with differential privacy applied to stochastic gradient descent (DP-SGD) [1], the synthetic data is shown to improve the performance in multiple privacy-preserving cancer classification settings.
- **Chapter 6. Sharing Generative Models to Overcome Data Scarcity Issues in Medical Imaging.** Generative models are proposed and validated as an alternative to private patient data, particularly due to their arguably less restrictive data sharing constraints. Furthermore, the *medigan* library is introduced as a platform for sharing such pretrained generative models in medical imaging. *medigan* enables researchers to generate high-quality synthetic data across diverse imaging modalities, which, hence, not necessarily require direct access to sensitive patient data. Additionally, comprehensive experimental evaluation is performed to benchmark and compare the generative models in the *medigan* library based on the FID metric. To this end, apart from using FID's vanilla feature extractors based on natural images [50], the usage of medical imaging domain-specific feature extractors based on radiology images [151] is proposed and analyzed.

1.6. Contributions and Document Structure

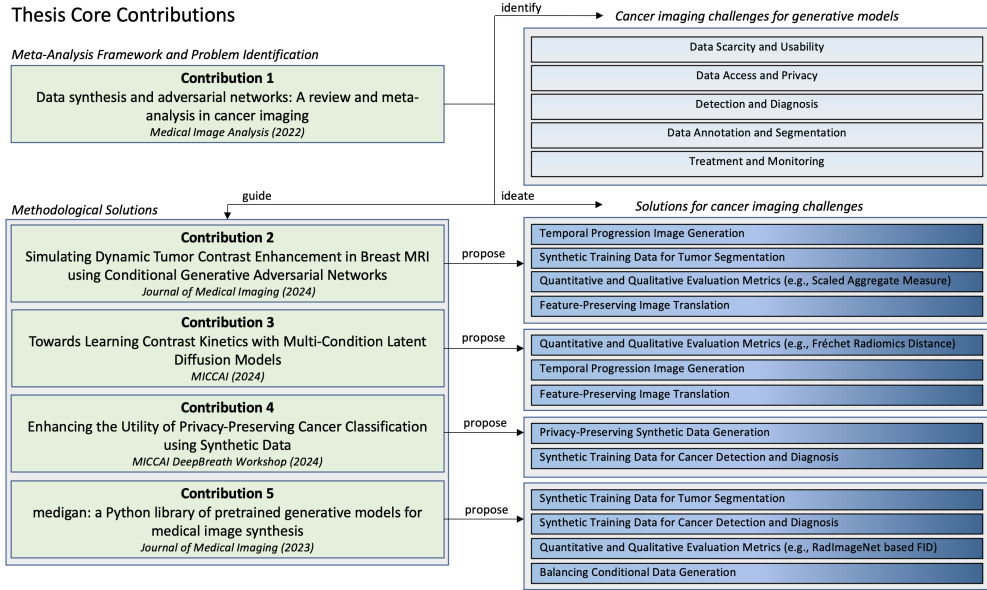


Figure 1.1: Visualization of the structure of contents of the present thesis. An in-depth survey in Section 2 identifies current challenges and problems in cancer imaging and proposes approaches to address these problems based on generative modeling, while also introducing a novel meta-analysis framework to assess the trustworthiness of studies applying these approaches. The subsequent sections 3 - 6 present and evaluate various generative modeling methods and applications to define, analyze and address specific problems identified in Section 2.

Chapter 2

Cancer Image Synthesis: Challenges, Solutions and Their Validation

Generative models have been seen notable adoption in the field of cancer imaging. However, it remains an open question which, and to what extent, the solutions proposed in the literature can solve the prevailing problems in the field of cancer imaging. In this chapter, we assess this aspect, alongside the technical maturity and trustworthiness of the proposed methods. To this end, we provide a trustworthiness meta-analysis framework and conduct a comprehensive review of the literature uncovering the problems in cancer imaging as well as surveying 164 papers containing respective potential solutions to address these challenges using image synthesis and adversarial neural networks methodologies. For instance, such challenges include intra- and inter-observer variability, class imbalance, dataset shifts, and tumor heterogeneity, which complicate the detection, diagnosis, and treatment of cancer.

This chapter further guides the definition of the generative models presented in subsequent chapters of the present doctoral thesis, as these models and methods strive to address several of the cancer imaging challenges uncovered in this chapter.

2.1 Contribution: Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging

Summary:

The paper is structured as a comprehensive review that explores both the current challenges and potential solutions in cancer imaging using GANs and adversarial learning. The challenges are grouped into five categories:

1. **Data scarcity and imbalance:** GANs can generate synthetic data to overcome the limited availability of diverse, high-quality medical imaging datasets.
2. **Data access and privacy:** The paper addresses issues related to patient privacy and security, showing how GANs can create synthetic data to bypass direct access to sensitive patient data.
3. **Data annotation and segmentation:** The difficulties associated with manual annotation and the variability in segmentation are examined, with GANs proposed as a solution to automate these processes.
4. **Cancer detection and diagnosis:** GANs are applied to assist in early cancer detection, improving diagnostic accuracy by synthesizing realistic images.
5. **Treatment planning and monitoring:** The role of GANs in predicting treatment outcomes and generating imaging data to track tumor progression is discussed.

Further, the SynTRUST meta-analysis framework is presented as a critical tool for evaluating the robustness, reproducibility, and scalability of GAN-based solutions in medical imaging. This paper systematically reviews these challenges, maps relevant solutions using GANs, and provides a direction for future research in this field.

Main Contributions:

- **Comprehensive Review:** The paper presents a first thorough review of generative models and adversarial learning in cancer imaging, while also categorizing and mapping image synthesis solutions to the prevailing challenges in cancer image analysis.

- **Introduction of SynTRUST Framework:** The paper introduces the Synthesis Study Trustworthiness Test (SynTRUST), a novel framework designed to assess the reliability and validity of medical image synthesis studies.
- **Meta-analysis of Studies:** The paper meta-analyzes 16 promising studies in cancer imaging, applying the SynTRUST framework to evaluate the validation rigor and technical maturity of their proposed solutions.
- **Research Directions:** It highlights underexplored areas where generative models could have a significant impact in cancer imaging, offering guidance for future research.
- **Bridge Between Clinical and AI Research:** The paper strives to bridge the gap between clinical needs and AI research in cancer imaging, proposing solutions that address both medical and technological challenges.

Contributions | The dissertation author led the conceptualization, and development of the review as well as the SynTrust meta-analysis, including the ideation, design, validation, and manuscript preparation. Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., and Lekadir, K. provided critical feedback and assisted in manuscript preparation.

Declaration of Doctoral Use | None of the co-authors have used, either implicitly or explicitly, this publication for the development of their own doctoral thesis, nor do they intend to do so in the future.

Publication | Journal: Medical Image Analysis, ISSN: 1361-8415, Publisher: Elsevier, Journal Citation Reports (JCR) 2023, Impact Factor: 10.7

Copyright Statement | This work was originally published as: "Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., Lekadir, K. (2023). Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Medical Image Analysis*, 84, 102704."

<https://doi.org/10.1016/j.media.2022.102704>



Survey paper

Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging

Richard Osuala^{a,*}, Kaisar Kushibar^a, Lidia Garrucho^a, Akis Linardos^a, Zuzanna Szafranowska^a, Stefan Klein^b, Ben Glocker^c, Oliver Diaz^{a,1}, Karim Lekadir^{a,1}

^a Artificial Intelligence in Medicine Lab (BCN-AIM), Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

^b Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands

^c Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK

ARTICLE INFO

Keywords:

Generative Adversarial Network
Adversarial training
Synthetic data
Trustworthiness

ABSTRACT

Despite technological and medical advances, the detection, interpretation, and treatment of cancer based on imaging data continue to pose significant challenges. These include inter-observer variability, class imbalance, dataset shifts, inter- and intra-tumour heterogeneity, malignancy determination, and treatment effect uncertainty. Given the recent advancements in image synthesis, Generative Adversarial Networks (GANs), and adversarial training, we assess the potential of these technologies to address a number of key challenges of cancer imaging. We categorise these challenges into (a) data scarcity and imbalance, (b) data access and privacy, (c) data annotation and segmentation, (d) cancer detection and diagnosis, and (e) tumour profiling, treatment planning and monitoring. Based on our analysis of 164 publications that apply adversarial training techniques in the context of cancer imaging, we highlight multiple underexplored solutions with research potential. We further contribute the Synthesis Study Trustworthiness Test (*SynTRUST*), a meta-analysis framework for assessing the validation rigour of medical image synthesis studies. *SynTRUST* is based on 26 concrete measures of thoroughness, reproducibility, usefulness, scalability, and tenability. Based on *SynTRUST*, we analyse 16 of the most promising cancer imaging challenge solutions and observe a high validation rigour in general, but also several desirable improvements. With this work, we strive to bridge the gap between the needs of the clinical cancer imaging community and the current and prospective research on data synthesis and adversarial networks in the artificial intelligence community.

1. Introduction

1.1. The burden of cancer and early detection

The evident improvement in global cancer survival in the last decades is arguably attributable not only to health care reforms, but also to advances in clinical research (e.g., targeted therapy based on molecular markers) and diagnostic imaging technology e.g whole-body magnetic resonance imaging (MRI) (Messiou et al., 2019), and positron emission tomography-computed tomography (PET-CT) (Arnold et al., 2019). Nonetheless, cancers still figure among the leading causes of morbidity and mortality worldwide (Ferlay et al., 2015), with an approximated 9.6 million cancer related deaths in 2018 (World Health Organization, 2018). The most frequent cases of cancer death worldwide

in 2018 are lung (1.76 million), colorectal (0.86 million), stomach (0.78 million), liver (0.78 million), and breast (0.63 million) (World Health Organization, 2018). These figures are prone to continue to increase in consequence of the ageing and growth of the world population (Jemal et al., 2011).

A large proportion of the global burden of cancer could be prevented due to treatment and early detection (Jemal et al., 2011). For example, an early detection can provide the possibility to treat a tumour before it acquires critical combinations of genetic alterations (e.g., metastasis with evasion of apoptosis Hanahan and Weinberg, 2000). Solid tumours become detectable by medical imaging modalities only at an approximate size of 10^9 cells ($\approx 1 \text{ cm}^3$) after evolving from a single neoplastic cell typically following a Gompertzian (Norton et al., 1976) growth pattern (Frangioni, 2008).² To detect and diagnose tumours, radiologists

* Corresponding author.

E-mail address: richard.osuala@ub.edu (R. Osuala).

¹ Authors contributed equally

² In vitro studies reported a *theoretical* detection limit around 10^5 to 10^6 for human cancer cell lines using PET. In clinical settings, the *theoretical* detection limit is larger and depends, among others, on background radiation, cancer cell line, and cancer type (Fischer et al., 2006).

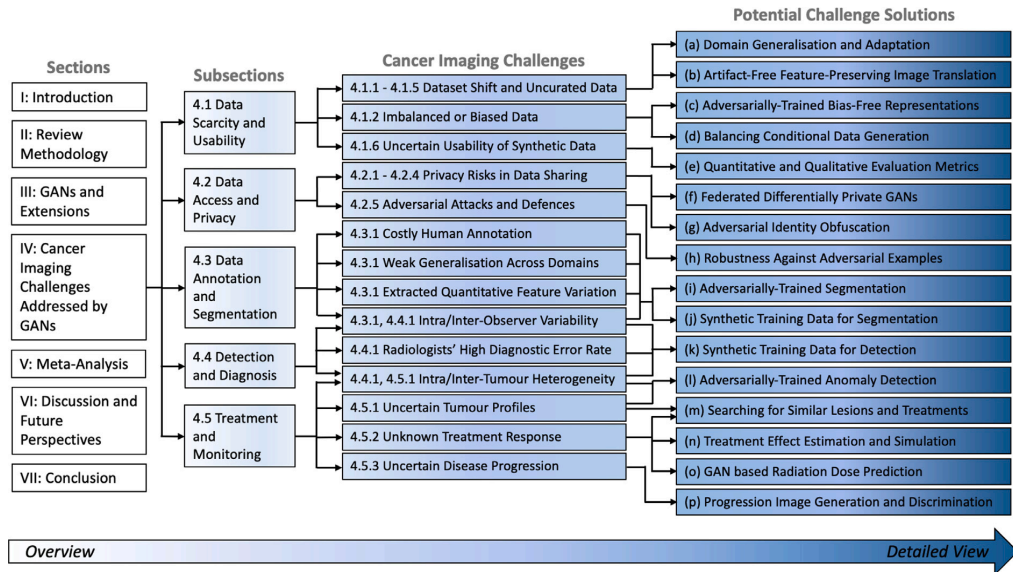


Fig. 1. Section organisation: Illustration of the structure of the present paper starting with paper sections on the left and going into detail towards the right culminating in a selection of cancer imaging challenges and solutions. These solutions (a)–(p) are part of the solutions found in the surveyed GAN literature or are proposed extensions thereof, as is further discussed in Section 4.

inspect, normally by visual assessment, medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound (US), X-ray mammography (MMG), PET (Frangioni, 2008; Itri et al., 2018; McCreddie and Oliver, 2009).

Medical imaging data evaluation is time demanding and therefore costly in nature. In addition, volumes of new technologies (e.g., digital breast tomosynthesis Swiecicki et al., 2021) become available and studies generally show an extensive increase in analysable imaging volumes (McDonald et al., 2015). Also, the diagnostic quality in radiology varies and is very much dependent on the personal experience, skills and invested time of the data examiner (Itri et al., 2018; Elmore et al., 1994; Woo et al., 2020). Hence, to decrease cost and increase quality, automated or semi-automated diagnostic tools can be used to assist radiologists in the decision-making process. Such diagnostic tools comprise traditional machine learning, but also recent deep learning methods, which promise an immense potential for detection performance improvement in radiology.

1.2. The promise of deep learning and the need for data

The rapid increase in graphics processing unit (GPU) processing power has allowed training deep learning algorithms such as convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1989, 1998) on large image datasets achieving impressive results in Computer Vision (CireAan et al., 2012; Krizhevsky et al., 2012), and Cancer Imaging (Cireşan et al., 2013). In particular, the success of AlexNet in the 2012 ImageNet challenge (Krizhevsky et al., 2012) triggered an increased adoption of deep neural networks to a multitude of problems in numerous fields and domains including medical imaging, as reviewed in Shen et al. (2017), Zhou et al. (2021) and Litjens et al. (2017). Despite the increased use of medical imaging in clinical practice, the public availability of medical imaging data remains limited (McDonald et al., 2015). This represents a key impediment for the training, research, and use of deep learning algorithms in radiology and oncology. Clinical centres refrain from sharing such data for ethical, legal, technical, and financial (e.g., costly annotation) reasons (Bi et al., 2019).

Such cancer imaging data not only is necessary to train deep learning models, but also to provide them with sufficient learning possibility to acquire robustness and generalisation capabilities. We define robustness as the property of a predictive model to remain accurate despite of variations in the input data (e.g., noise levels, resolution, contrast, etc.). We refer to a model's generalisation capability as its property of preserving predictive accuracy on new data from unseen sites, hospitals, scanners, etc. Both of these properties are in particular desirable in cancer imaging considering the frequent presence of biased or unbalanced data with sparse or noisy labels.³ Both robustness and generalisation are essential to demonstrate the trustworthiness of a deep learning model for usage in a clinical setting, where every edge-case needs to be detected and a false negative can potentially cost the life of a patient.

1.3. Synthetic cancer imaging data

We hypothesise that the variety of data needed to train robust and well-generalising deep learning models for cancer images can be largely synthetically generated using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). The adversarial learning scheme in GANs is based on a generator that generates synthetic (alias 'fake') samples of a target distribution trying to fool a discriminator, which classifies these samples as either real or fake. Various papers have provided reviews of GANs in the medical imaging domain (Yi et al., 2019; Kazeminia et al., 2020; Tschuchnig et al., 2020; Sorin et al., 2020; Lan et al., 2020; Singh and Raza, 2020), but they focused on general presentation of the main methods and possible applications. In cancer imaging, however, there are specificities and challenges that call for specific implementations and solutions based on GANs and the adversarial learning scheme at large, including:

- (i) the small size and complexity of cancerous lesions

³ Alongside tumour manifestation heterogeneity, and multi-centre, multi-organ, multi-modality, multi-scanner, and multi-vendor data.

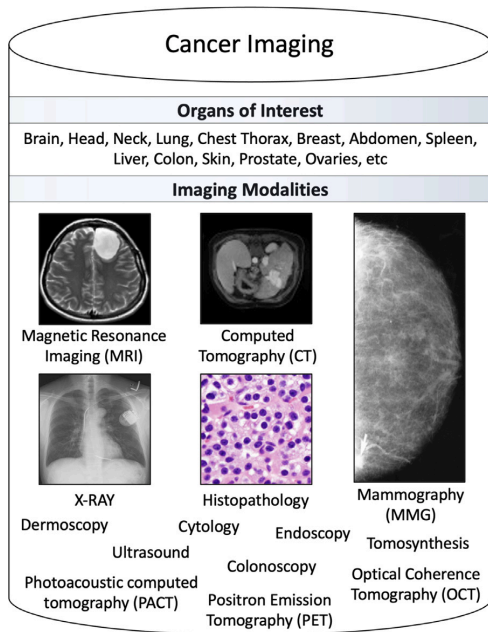


Fig. 2. Overview of the most common organs and modalities targeted by the surveyed cancer imaging publications. A respective histogram that shows the number of papers per modality and per organ can be found in Fig. 15.

- (ii) the high heterogeneity between tumours within as well as between patients and cancer types
- (iii) the difficulty to annotate, delineate and label cancer imaging studies at large scale
- (iv) the high data imbalance in particular between healthy and pathological subjects or between benign and malignant cases
- (v) the difficulty to gather large consented datasets from highly vulnerable patients undergoing demanding care plans

Hence, the present paper contributes a unique perspective and comprehensive analysis of adversarial networks attempting to address the specific challenges in the cancer imaging domain. To the authors' best knowledge, this is the first survey that exclusively focuses on GANs and adversarial training in cancer imaging. In this context, we define cancer imaging as the entirety of approaches for research, diagnosis, and treatment of cancer based on medical images. Our survey comprehensively analyses cancer imaging GAN and adversarial training applications focusing on radiology modalities. As presented in Fig. 2, we recognise that non-radiology modalities are also widely used in cancer imaging. For this reason, we do not restrict the scope of our survey to radiology, but rather also analyse relevant publications in these other modalities including histopathology and cytopathology (e.g., in Section 4.5), and dermatology (e.g., in Sections 4.3 and 4.4).

Further, our survey uncovers and highlights promising research directions for adversarial networks and image synthesis that can facilitate the sustainable adoption of AI in clinical oncology and radiology.

1.4. Section organisation

The remainder of this paper is organised as follows. In Section 2, we introduce the methodology of this review. Section 3 provides an overview of GANs and highlights extensions of the adversarial learning framework relevant to cancer imaging. Section 4 contains the main

contribution that encompasses the systematic review of challenges of cancer imaging and potential solutions based on adversarial networks. This organisation is depicted in more detail in Fig. 1.

The different challenges are categorised into groups in the Sections 4.1, 4.2, 4.3, 4.4, and 4.5. Each of the challenges categories contains several specific cancer imaging challenges, which we introduce and discuss in 4.1.1–4.5.3. The sections are organised in an independent way allowing the reader to directly jump to a particular cancer imaging category (4.1–4.5) of interest without requiring context from previous sections. For each of the specific challenges, we survey and discuss potential solutions, as depicted in Fig. 1(a)–(p).

The subsequent Section 5 contains our second core contribution, which consists of the *SynTRUST* framework for systematic analysis of trustworthiness criteria of image synthesis and adversarial training publications in medical imaging. Based on this framework, we meta-analyse a set of studies selected based on their strong performance and promising methodology for solving a specific cancer imaging challenge.

After learning how and to what extent image synthesis and adversarial training solutions have addressed cancer imaging challenges in the past, we highlight and discuss prospective avenues of future research in the Discussion Section 6 and point out unexploited potential of image synthesis and adversarial networks in cancer imaging.

2. Review methodology

Our review comprises two comprehensive literature screening processes. The first screening process surveyed the current challenges in the field of cancer imaging with a focus on radiology imaging modalities. After screening and gaining a deepened understanding of AI-specific and general cancer imaging challenges, we grouped these challenges for further analysis into the following five categories.

- *Data scarcity and usability challenges* (Section 4.1); discussing dataset shifts, class imbalance, fairness, generalisation, domain adaptation and the evaluation of synthetic data.
- *Data access and privacy challenges* (Section 4.2); comprising patient data sharing under privacy constraints, security risks, and adversarial attacks.
- *Data annotation and segmentation challenges* (Section 4.3); discussing costly human annotation, high inter and intra-observer variability, and the consistency of extracted quantitative features.
- *Detection and diagnosis challenges* (Section 4.4); analysing the challenges of high diagnostic error rates among radiologists, early detection, and detection model robustness.
- *Treatment and monitoring challenges* (Section 4.5); examining challenges of high inter and intra-tumour heterogeneity, phenotype to genotype mapping, treatment effect estimation and disease progression.

The second screening process comprised first of a generic and second a specific literature search to find all papers that apply adversarial learning (i.e. GANs) to cancer imaging. In the generic literature search, generic search queries such as 'Cancer Imaging GAN', 'Tumour GANs' or 'Nodule Generative Adversarial Networks' were used to recall a high number of papers. The specific search focused on answering key questions of interest to the aforesaid challenges such as 'Carcinoma Domain Adaptation Adversarial', 'Skin Melanoma Detection GAN', 'Brain Glioma Segmentation GAN', or 'Cancer Treatment Planning GAN'.

In Section 4, we map the papers that propose adversarial training and GAN applications applied to cancer imaging (second screening) to the surveyed cancer imaging challenges (first screening). The mapping of these GAN-related papers to challenge categories facilitates analysing the extent to which existing solutions solve the current cancer imaging challenges and helps to identify gaps and further potential for adversarial networks in this field. The mapping is based on the evaluation criteria used in the GAN-related papers and on the relevance

of the reported results to the corresponding section. For example, if a GAN generates synthetic data that is used to train and improve a tumour detection model, then this paper is assigned to the detection and diagnosis challenge Section 4.4. If a paper describes a GAN that improves a segmentation model, then this paper is assigned to the segmentation and annotation challenge Section 4.3, and so forth.

To gather the literature (e.g., first papers describing cancer imaging challenges, second papers proposing GAN solutions), we have searched in medical imaging, computer science and clinical conference proceedings and journals, but also freely on the web using the search engines Google, Google Scholar, and PubMed. After retrieving all papers with a title related to the subject, their abstract was read to filter out non-relevant papers. A full-text analysis was done for the remaining papers to determine whether they were to be included into our manuscript. We analysed the reference sections of the included papers to find additional relevant literature, which also underwent filtering and full-text screening. Applying this screening process, we reviewed and included a total of 164 GAN and adversarial training cancer imaging publications comprising both peer-reviewed articles and conference papers, but also relevant preprints from arXiv and bioRxiv.

Details about these 164 cancer imaging applications can be found in Tables 2–6. The distribution of these publications across challenge category, year, modality, and anatomy is outlined in Fig. 15.

The methodology for deriving and applying the *SynTRUST* meta-analysis framework, which assesses the validity and trustworthiness of medical image synthesis studies, is provided in Section 5.

3. GANs and extensions

3.1. Introducing the theoretical underpinnings of GANs

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are a type of generative model with a differentiable generator network (Goodfellow et al., 2016). GANs are formalised as a minimax two-player game, where the generator network (G) competes against an adversary network called discriminator (D). As visualised in Fig. 3, given a random noise distribution z , G generates samples $x = G(z; \theta(g))$ that D classifies as either real (drawn from training data, i.e. $x \sim p_{data}$) or fake (drawn from G, i.e. $x \sim p_g$). x is either sampled from p_{data} or from p_g with a probability of 50%. D outputs a value $p = D(x; \theta(d))$ indicating the probability that x is a real training example rather than one of G's fake samples (Goodfellow et al., 2016). As defined by Goodfellow et al. (2014), the task of the discriminator can be characterised as binary classification (CLF) of samples x . Hence, the discriminator can be trained using binary-cross entropy resulting in the following loss function L_D :

$$L_D = -\mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (1)$$

D's training objective is to minimise L_D (or maximise $-L_D$) while the goal of the generator is the opposite (i.e. minimise $-L_D$) resulting in the value function $V(G, D)$ of a two-player zero-sum game between D and G:

$$\min_G \max_D V(G, D) = \min_G \max_D [\mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (2)$$

In theory, in convergence, the generator's samples become indistinguishable from the real training data ($x \sim p_{data} = p_g$) and the discriminator outputs $p = 1/2$ for any given sample x (Goodfellow et al., 2016). As this is a state where both D and G cannot improve further on their objective by changing only their own strategy, it represents a Nash equilibrium (Farnia and Ozdaglar, 2020; Nash et al., 1950). In practice, achieving convergence for this or related adversarial training schemes is an open research problem (Kodali et al., 2017; Mescheder et al., 2018; Farnia and Ozdaglar, 2020).

3.2. Extensions of the Vanilla GAN methodology

As indicated by Fig. 4, numerous extensions of GANs have shown to generate synthetic images with high realism (Karras et al., 2017, 2019, 2020; Chan et al., 2020) and under flexible conditions (Mirza and Osindero, 2014; Odena et al., 2017; Park et al., 2018). GANs have been successfully applied to generate high-dimensional data such as images and, more recently, have also been proposed to generate discrete data (Hjelm et al., 2017). Apart from image generation, GANs have also widely been proposed and applied for paired and unpaired image-to-image translation, domain-adaptation, data augmentation, image inpainting, image perturbation, super-resolution, and image registration and reconstruction (Yi et al., 2019; Kazeminia et al., 2020; Wang et al., 2019b).

Table 1 introduces a selection of common GAN extensions found to be frequently applied to cancer imaging. For each GAN methodology in this and the Tables 1–6, we define the 'Task' describing the application of the respective adversarial network. For instance, in 'noise-to-image synthesis' the input into the generator G consists of a noise vector that G translates into an image. A further input into G can be a class label as in 'class-conditional-image-synthesis' based on which an output is generated that corresponds to this class. Paired and unpaired translation refer to the task where the input into G is a sample (e.g. an image in the source domain) based on which G generates another sample (e.g. an image in the target domain). This translation is paired if the training data consists of target and source domain sample pairs. The key characteristics of each of the GAN extensions of Table 1 are described in the following paragraphs.

3.2.1. Noise-to-image GAN extensions

As depicted in blue in Fig. 3, cGAN adds a discrete label as conditional information to the original GAN architecture that is provided as input to both generator and discriminator to generate class conditional samples (Mirza and Osindero, 2014).

AC-GAN feeds the class label only to the generator while the discriminator is tasked with correctly classifying both the class label and whether the supplied image is real or fake (Odena et al., 2017).

WGAN is motivated by mathematical rationale and based on the Wasserstein-1 distance (alias 'earth mover distance' or 'Kantorovich distance') between two distributions. WGAN extends on the theoretic formalisation and optimisation objective of the vanilla GAN to better approximate the distribution of the real data. By applying an alternative loss function (i.e. Wasserstein loss), the discriminator (alias 'critic' or ' f_w ') maximises – and the generator minimises – the difference between the critic's scores for generated and real samples. A important benefit of WGAN is the empirically observed correlation of the loss with sample quality, which helps to interpret WGAN training progress and convergence (Arjovsky et al., 2017).

In WGAN, the weights of the critic are clipped, which means they have to lie within a compact space $[-c, c]$. This is needed to fulfil that the critic is constraint to be in the space of 1-Lipschitz functions. With clipped weights, however, the critic is biased towards learning simpler functions and prone to have exploding or vanishing gradients if the clipping threshold c is not tuned with care (Gulrajani et al., 2017; Arjovsky et al., 2017).

In WGAN-GP, the weight clipping constraint is replaced with a gradient penalty. Gradient penalty of the critic is a tractable and soft version of the following notion: By constraining that the norm of the gradients (i.e. the critic) would fulfil the 1-Lipschitz criterion without the need of weight clipping. Compared, among others, to WGAN, WGAN-GP was shown to have improved training stability (i.e. across many different GAN architectures), training speed, and sample quality (Gulrajani et al., 2017).

DCGAN generates realistic samples using a convolutional network architecture with batch normalisation (Ioffe and Szegedy, 2015) for

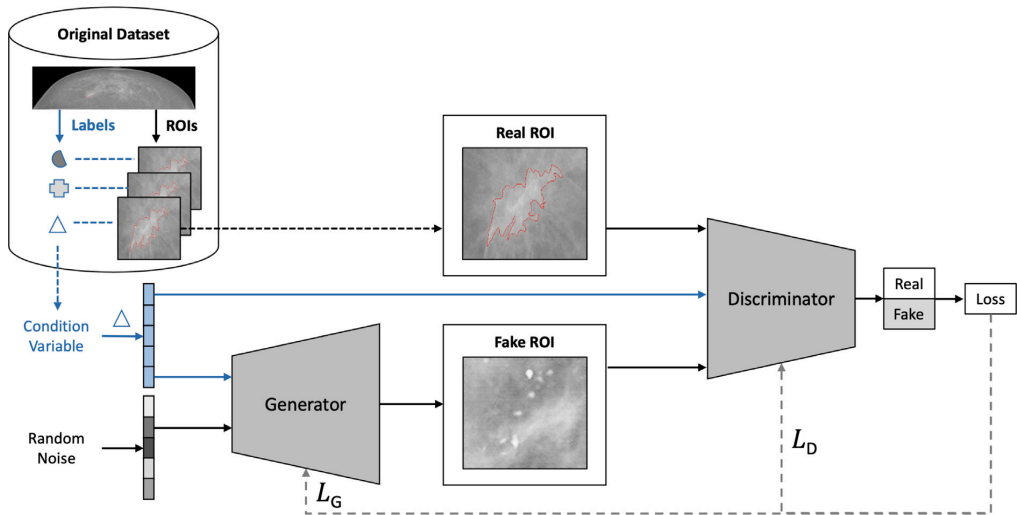


Fig. 3. An example of a generic GAN architecture applied to generation of synthetic mammography region of interest (ROI) images based on the INbreast dataset (Moreira et al., 2012). Note that including the ‘Condition’ depicted in blue colour extends the vanilla GAN architecture (Goodfellow et al., 2014) to the cGAN architecture (Mirza and Osindero, 2014).

both generator and discriminator and progressively increases the spatial dimension in the layers of the generator using transposed convolution (alias ‘fractionally-strided convolution’) (Radford et al., 2015).

PGGAN is tested with loss and configurations introduced in WGAN GP. It starts by generating low pixel resolution images, but progressively adds new layers to the generator and discriminator during training resulting in increased pixel resolution and finer image details. It is suggested that after early convergence of initial low-resolution layers, the introduced additional layers enforce the network to only refine the learned representations by increasingly smaller-scale effects and features (Karras et al., 2017).

In SRGAN, the generator transforms a low-resolution (LR) to a high-resolution (HR, alias ‘super-resolution’) image, while the discriminator learns to distinguish between real high-resolution images and fake super-resolution images. Apart from an adversarial loss, a perceptual loss called ‘content loss’ measures how well the generator represents higher level image features. This content loss is computed as the euclidean distance between feature representations of the reconstructed image and the reference image based on feature maps of a pretrained 19 layer VGG (Simonyan and Zisserman, 2014) network (Ledig et al., 2017).

3.2.2. Image-to-image GAN extensions

In image-to-image translation, a mapping is learned from one image distribution to another. For example, images from one domain can be transformed to resemble images from another domain via a mapping function implemented by a GAN generator.

CycleGAN achieves realistic unpaired image-to-image translation using two generators (G , F) with one traditional adversarial loss each and an additional cycle-consistency loss. Unpaired image-to-image translation transforms images from domain X to another domain Y in the absence of paired training data i.e. corresponding image pairs for both domains. In CycleGAN, the input image x from domain X is translated by generator $G(x)$ to resemble a sample from domain Y . Next, the sample is translated back from domain Y to domain X by generator $F(G(x))$. The cycle consistency loss enforces that $F(G(x)) \approx x$ (forward cycle consistency) and that $G(F(y)) \approx y$ (backward cycle consistency) (Zhu et al., 2017).

Both pix2pix and SPADE are used in paired image-to-image translation where corresponding image pairs for both domains X and Y

are available. pix2pix (alias ‘condGAN’) is a conditional adversarial network that adapts the U-Net architecture⁴ (Ronneberger et al., 2015) for the generator to facilitate encoding an conditional input image into a latent representation before decoding it back into an output image. pix2pix uses L1 loss to enforce low level (alias ‘low frequency’) image reconstruction and a patch-based discriminator (‘PatchGAN’) to enforce high level (alias ‘high frequency’) image reconstruction that the authors suggest to interpret as texture/style loss. Note that the input into the PatchGAN discriminator is a concatenation⁵ of the original image (i.e. the generator’s input image; e.g. this can be a segmentation map) and the real/generated image (i.e. the generator’s output image) (Isola et al., 2017).

In SPADE, the generator architecture does not rely on an encoder for downsampling, but uses a conditional normalisation method during upsampling instead: A segmentation mask as conditional input into the SPADE generator is provided to each of its upsampling layers via spatially-adaptive residual blocks. These blocks embed the masks and apply two two-layer convolutions to the embedded mask to get two tensors with spatial dimensions. These two tensors are multiplied/added to each upsampling layer prior to its activation function. The authors demonstrate that this type of normalisation achieves better fidelity and preservation of semantic information in comparison to other normalisation methods that are commonly applied in neural networks (e.g., Batch Normalisation). The multi-scale discriminators and the loss functions from pix2pixHD (Wang et al., 2018a) are adapted in SPADE, which contains a hinge loss (i.e. as substitute of the adversarial loss), a perceptual loss, and a feature matching loss (Park et al., 2019).

3.2.3. GAN network architectures and adversarial loss

For further methodological detail on the aforementioned GAN methods, loss functions, and architectures, we point the interested reader to the GAN methods review by Wang et al. (2019b). Due to the

⁴ To reduce information loss in latent space compression, U-Net uses skip connections between corresponding layers (e.g., first to last) in the encoder and decoder.

⁵ Note the concatenation of real_A and fake_B before computing the loss in the discriminator backward pass (L93) in the authors’ pix2piximplementation.

Table 1
A selection of the GAN architectures that we found to be the ones most frequently applied in cancer imaging.

Publication	Input G	Input D	Losses	Task
Noise to image				
GAN (Goodfellow et al., 2014)	Noise	Image	Binary cross-entropy based adversarial loss (L_{ADV})	Noise-to-image synthesis
Conditional GAN (cGAN) (Mirza and Osindero, 2014)	Noise & label	Image & label	L_{ADV}	Class-conditional image synthesis
Auxiliary classifier GAN (AC-GAN) (Odena et al., 2017)	Noise & label	Image	L_{ADV} & cross-entropy loss (label classification)	Class-conditional image synthesis
Deep convolutional GAN (DCGAN) (Radford et al., 2015)	Noise	Image	L_{ADV}	Noise-to-image synthesis
Wasserstein GAN (WGAN) (Arjovsky et al., 2017)	Noise	Image	Wasserstein loss (L_{WGAN})	Noise-to-image synthesis
WGAN gradient penalty (WGAN GP) (Gulrajani et al., 2017)	Noise	Image	L_{WGAN} with GP ($L_{WGAN-GP}$)	Noise-to-image synthesis
Progressively growing GAN (PGGAN) (Karras et al., 2017)	Noise	Image	$L_{WGAN-GP}$	Noise-to-image synthesis
Image to Image				
Super-Resolution GAN (SRGAN) (Ledig et al., 2017)	Image (LR)	Image (HR)	L_{ADV} & content loss (based on VGG features)	Super-resolution
CycleGAN (Zhu et al., 2017)	Source image	Target image	L_{ADV} & cycle consistency loss & identity loss	Unpaired image-to-image translation
pix2pix (Isola et al., 2017)	Source image	Concatenated source and target images	L_{ADV} & reconstruction loss (i.e. L1)	Paired image-to-image translation
SPAtially-adaptive (DE)normalization (SPADE) (Park et al., 2019)	Noise or encoded source image & segmentation map	Concatenated target image and segmentation map	Hinge & perceptual & feature matching losses (from Wang et al., 2018a)	Paired image-to-image translation

image processing capabilities of CNNs (LeCun et al., 1989), the above-mentioned GAN architectures generally rely on CNN layers internally. Recently, TransGAN (Jiang et al., 2021) and VQGAN (Esser et al., 2021) were proposed, which diverges from the CNN design pattern to using Transformer Neural Networks (Vaswani et al., 2017). Due to the promising performances of these approaches in computer vision tasks, we encourage future studies to investigate the potential of transformer-based GANs for applications in medical and cancer imaging.

Multiple deep learning architectures apply the adversarial loss proposed in Goodfellow et al. (2014) together with other loss functions (e.g., segmentation loss functions) for other tasks than image generation (e.g., image segmentation). This adversarial loss is useful for unsupervised learning of features and representations that are invariant to some part of the training data. For instance, adversarial learning can be useful to discriminate a domain to learn domain-invariant representations (Ganin and Lempitsky, 2015), as has been successfully demonstrated for medical images (Kamnitsas et al., 2017). Such methods that apply the adversarial loss internally are referred to as ‘adversarial training’ methods and are included in the scope of our survey. That is, we include and consider all relevant cancer imaging papers that apply or build upon the adversarial learning scheme defined in Goodfellow et al. (2014), which comprises GANs as well as adversarial training methods.

4. Cancer imaging challenges addressed by data synthesis and adversarial networks

In this section we follow the structure presented in Fig. 1, where we categorise cancer imaging challenges into five categories consisting of *data scarcity and usability* (4.1), *data access and privacy* (4.2), *data annotation and segmentation* (4.3), *detection and diagnosis* (4.4), and *treatment and monitoring* (4.5). In each subsection, we group and analyse respective cancer imaging challenges and discuss the potential and the limitations of corresponding GAN-based data synthesis and adversarial training solutions. In this regard, we also identify and highlight key needs to be addressed by researchers in the field of cancer imaging

GANs towards solving the surveyed cancer imaging challenges. We provide respective Tables 2–6 for each Sections 4.1–4.5 containing relevant information (publication, method, dataset, modality, task, highlights) for all of the reviewed cancer imaging GAN solutions.

Chronology of key innovations. The most commonly applied adversarial network methodologies in cancer imaging are summarised chronologically in Fig. 4. Next to each network (a)–(m), the number of occurrence per cancer imaging challenge category 4.1–4.5 is highlighted.

Following Vanilla GANs 4(a), four main lines of innovations have been widely adopted in cancer imaging. These are methods that condition the synthetic data generation e.g. cGAN 4(b), methods that improve upon the network architecture e.g. DCGAN 4(c), methods that improve upon the adversarial loss function e.g. WGAN 4(g), and methods that backpropagate the adversarial loss for representation learning, e.g. domain-invariant representations 4(d).

As to conditional methods, further key innovations have been AC-GAN’s 4(f) discriminator classifying the input condition, and methods that conditioning the generation based on an input image using additional reconstruction (e.g., pix2pix 4(e), cycleGAN 4(h)) or perceptual (e.g., SRGAN 4(i)) losses. Recent approaches (e.g., SPADE 4(l)) innovate regarding how the input image is provided to the generator network, e.g., via spatially-adaptive residual blocks in upsampling layers.

WGAN’s 4(g) loss based on the discriminator estimating the Wasserstein-1 distance between real and synthetic image distributions is a widely used and extended (e.g., WGAN-GP 4(j)) alternative to the vanilla binary-cross entropy adversarial loss in cancer imaging.

The architectural innovation of progressive network growing 4(k) unlocked high-resolution cancer image generation and is adopted by recent approaches such as StyleGAN 4(m), which introduced adaptive instance normalisation and pioneered noise (and style condition) input via intermediate activation maps.

4.1. Data scarcity and usability challenges

4.1.1. Challenging dataset sizes and shifts

Although data repositories such as The Cancer Imaging Archive (TCIA) (Clark et al., 2013) have made a wealth of cancer imaging

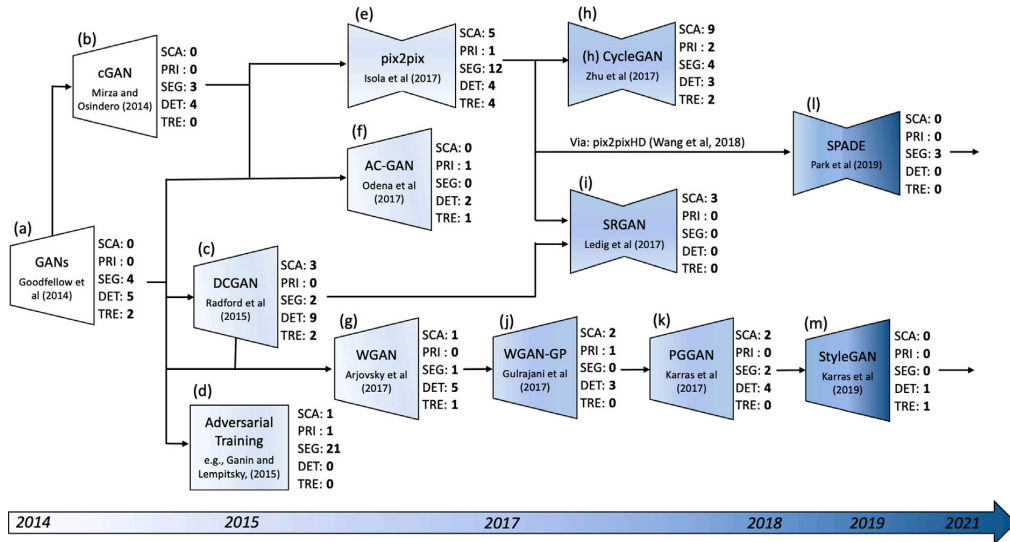


Fig. 4. The chronological evolution of adversarial networks in cancer imaging based on the commonly applied GAN types (a)–(m). The shape of the respective GAN indicates its mapping function, e.g., a trapezium represents a noise-to-image mapping function. For each cancer imaging challenge category surveyed in 4, the number of occurrences of each GAN type (a)–(m) is counted across reviewed publications. The illustration abbreviates challenge categories as SCA: 4.1 Data scarcity and usability challenges, PRI: 4.2 Data access and privacy challenges, SEG: 4.3 Data annotation and segmentation challenges, DET: 4.4 Detection and diagnosis challenges, TRE: 4.5 Treatment and monitoring challenges.

data available for research, the demand is still far from satisfied. As a result, data augmentation techniques are widely used to artificially enlarge the existing datasets, traditionally including simple spatial (e.g., flipping, rotation) or intensity transformations (e.g., noise insertion) of the true data. GANs have shown promise as a more advanced augmentation technique and have already seen use in medical and cancer imaging (Han et al., 2018; Yi et al., 2019).

Aside from the issue of lacking sizeable data, data scarcity often forces studies to be constrained on small-scale single-centre datasets. The resulting findings and models are likely to not generalise well due to diverging distributions between the (synthetic) datasets seen in training and those seen in testing or after deployment, a phenomenon known as dataset shift (Quionero-Candela et al., 2009).⁶ An example of this in clinical practice are cases where training data is preselected from specific patient sub-populations (e.g., only high-risk patients) resulting in bias and limited generalisability to the broad patient population (Troyanskaya et al., 2020; Bi et al., 2019).

From a causality perspective, dataset shift can be split into several distinct scenarios (Castro et al., 2020):

- *Population shift*, caused by differences in age, sex, ethnicities etc.
- *Acquisition shift*, caused by differences in scanners, resolution, contrast etc.
- *Annotation shift*, caused by differences in annotation policy, annotator experience, segmentation protocols etc.
- *Prevalence shift*, caused by differences in the disease prevalence in the population, often resulting from artificial sampling of data
- *Manifestation shift*, caused by differences in how the disease is manifested

GANs may inadvertently introduce such types of dataset shifts (e.g., due to mode collapse Goodfellow et al., 2014), but it has been shown that this shift can be studied, measured and avoided (Santurkar

et al., 2018; Arora et al., 2018). GANs can be a sophisticated tool for data augmentation or curation (Diaz et al., 2021) and by calibrating the type of shift introduced, they have the potential to turn it into an advantage, generating diverse training data that can help models generalise better to unseen target domains. The research line studying this problem is called *domain generalisation*, and it has presented promising results for harnessing adversarial models towards learning of domain-invariant features (Zhou et al., 2021). GANs and adversarial training have been used in various ways in this context, using multi-source data to generalise to unseen targets (Rahman et al., 2019; Li et al., 2018) or in unsupervised domain generalisation using adaptive data augmentation to append adversarial examples iteratively (Volpi et al., 2018). As indicated in Fig. 1(a), the domain generalisation research line has recently been further extended to cancer imaging (Lafarge et al., 2019; Chen et al., 2021).

In the following, further cancer imaging challenges in the realm of data scarcity and usability are described and related GAN solutions are referenced. Given these challenges and solutions, we derive a workflow for clinical adoption of (synthetic) cancer imaging data, which is illustrated in Fig. 5.

4.1.2. Imbalanced data and fairness

Apart from the rise of data-hungry deep learning solutions and the need to cover the different organs and data acquisition modalities, a major problem that arises from data scarcity is that of imbalance—i.e. the overrepresentation of a certain type of data over others (Bi et al., 2019). In its more common form, imbalance of diagnostic labels can hurt a model's specificity or sensitivity, as a prior bias from the data distribution may be learned. The Lung Screening Study (LSS) Feasibility Phase exemplifies the common class imbalance in cancer imaging data: 325 (20.5%) suspicious lung nodules were detected in the 1586 first low-dose CT screening, of which only 30 (1.89%) were lung cancers (Gohagan et al., 2004, 2005; NLST Research Team, 2011). This problem directly translates to multi-task classification (CLF), with imbalance between different types of cancer leading to worse sensitivity on the underrepresented categories (Yu et al., 2013). It is important to note that by solving the imbalance with augmentation techniques,

⁶ More concretely, this describes a case of covariate shift (Quionero-Candela et al., 2009; Shimodaira, 2000) defined by a change of distribution within the independent variables between two datasets.

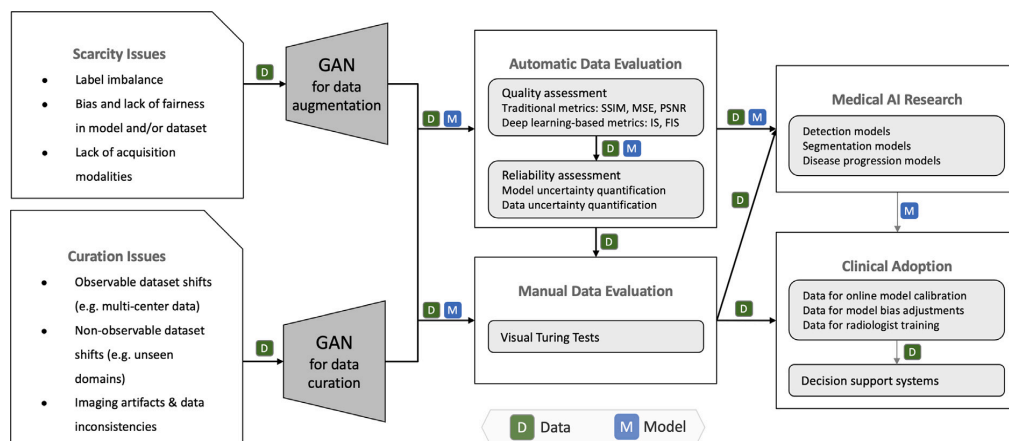


Fig. 5. Illustration of a workflow that applies GANs to the challenges of data scarcity and data curation. After the GAN generates synthetic data specific to the issue at hand, the data is automatically and manually evaluated before further used in medical AI research. Ultimately, both synthetic data and medical AI models are integrated as decision support tools into clinical practice.

bias is introduced as the prior distribution is manipulated, causing prevalence shift. As such, the test set should preserve the population statistics. Aside from imbalance of labels, more insidious forms of imbalance such as that of race/ethnicity (Adamson and Smith, 2018) or gender (Larrazabal et al., 2020) of patients are easily omitted in studies. This leads to fairness problems in real world applications as underrepresenting such categories in the training set will hurt performance on these categories in the real world (population shift) (Li et al., 2021a). Because of their potential to generate synthetic data, GANs are a promising solution to the aforementioned problems and have already been thoroughly explored in this regard in Computer Vision (Sampath et al., 2021; Mullick et al., 2019). Concretely, the discriminator and generator can be conditioned on underrepresented labels, forcing the generator to create images for a specific class,⁷ as indicated in Fig. 1(d). Many lesions classifiable by complex scoring systems such as RADS reporting are rare and, hence, effective conditional data augmentation is needed to improve the recognition of such lesions by ML detection models (Kazuhiro et al., 2018). GANs have already been used to adjust label distributions in imbalanced cancer imaging datasets, e.g. by generating underrepresented grades in a risk assessment scoring system (Hu et al., 2018b) for prostate cancer. A further promising applicable method is to enrich the data using a related domain as proxy input (Addepalli et al., 2020). Towards the goal of a more diverse distribution of data with respect to gender and ethnicity, similar principles can be applied. For instance, Li et al. (2021a) proposed an adversarial training scheme to improve fairness in classification of skin lesions for underrepresented groups (age, sex, skin tone) by learning a neutral representation using an adversarial bias discrimination loss. Fairness imposing GANs can also generate synthetic data with a preference for underrepresented groups, so that models may ingest a more balanced dataset, improving demographic parity without excluding data from the training pipeline. Such models have been trained in computer vision tasks (Sattigeri et al., 2018; Wang et al., 2019a; Zhang et al., 2018a; Xu et al., 2018; Beutel et al., 2017), but corresponding research on medical and cancer imaging denoted by Fig. 1(c) has been limited (Li et al., 2021a; Ghorbani et al., 2020).

⁷ The class can be something as simple as ‘malignant’ or ‘benign’, or a more complex score for risk assessment of a tumour such as the BiRADs scoring system for breast tumours (Lieberman and Menell, 2002).

4.1.3. Cross-modal data generation

In cancer, multiple acquisition modalities are enlisted in clinical practice (Kim et al., 2016; Chen et al., 2017; Barbaro et al., 2017; Chang et al., 2020b,a); thus automated diagnostic models should ideally learn to interpret various modalities as well or learn a shared representation of these modalities. Conditional GANs offer the possibility to generate one or multiple (Yurt et al., 2019; Li et al., 2019a; Zhou et al., 2020) modalities from another, alleviating the need to actually perform the potentially more harmful screenings—i.e. high-dose CT, PET—that expose patients to radiation, or require invasive contrast agents such as intravenous iodine-based contrast media (ICM) in CT (Haubold et al., 2021), gadolinium-based contrast agents in MRI (Zhao et al., 2020a) (in Table 5) or radioactive tracers in PET (Wang et al., 2018b; Zhao et al., 2020b). Furthermore, extending the acquisition modalities used in a given task would also enhance the performance and generalisability of AI models, allowing them to learn shared representations among these imaging modalities (Bi et al., 2019; Hosny et al., 2018). Towards this goal, multiple GAN domain-adaptation solutions have been proposed to generate CT using MRI (Wolterink et al., 2017; Kearney et al., 2020b; Tanner et al., 2018; Kaiser and Albarqouni, 2019; Nie et al., 2017; Kazemifar et al., 2020; Prokopenko et al., 2019), PET from MRI (Wang et al., 2018b), PET from CT (Ben-Cohen et al., 2017; Bi et al., 2017) (in Table 5), and CT from PET as in Armanious et al. (2020), where also GAN-based PET denoising and MR motion correction are demonstrated. If not indicated otherwise, these image-to-image translation studies are outlined in Table 2. Because of its complexity, clinical cancer diagnosis is based not only on imaging but also non-imaging data (genomic, molecular, clinical, radiological, demographic, etc.). In cases where this data is readily available, it can serve as conditional input to GANs towards the generation of images with the corresponding phenotype-genotype mapping, as is also elaborated in regard to tumour profiling for treatment in Section 4.5.1. A multimodal cGAN was recently developed, conditioned on both images and gene expression code (Xu et al., 2020); however, research along this line is otherwise limited.

4.1.4. Feature hallucinations in synthetic data

As displayed in Fig. 6 and denoted in Fig. 1(b), conditional GANs can unintentionally⁸ hallucinate non-existent artifacts into a patient image. This is particularly likely to occur in cross-modal data augmentation, especially but not exclusively if the underlying dataset is

⁸ Intentional feature injection or removal is discussed in 4.2.5.

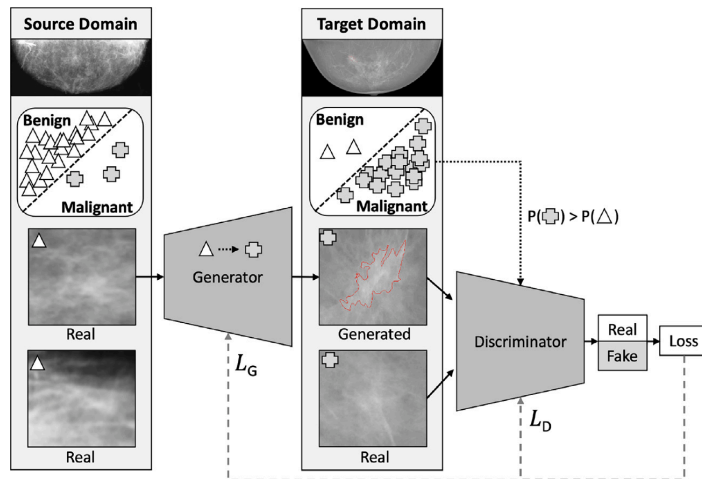


Fig. 6. Example of a GAN that translates Film Scanned MMG (source) to Full-Field Digital MMG (target). The generator transforms 'benign' source images (triangles) into 'malignant' target images (plus symbols). As opposed to source, the target domain contains more malignant MMGs than benign ones. If the discriminator thus learns to associate malignancy with realness, this incentivises the generator to inject malignant features (depicted by dotted arrows). For simplicity additional losses (e.g., reconstruction losses) are omitted.

imbalanced. For instance, Cohen et al. (2018a) describe GAN image feature hallucinations embodied by added and removed brain tumours in cranial MRI. The authors tested the relationship between the ratio of tumour images in the GAN target distribution and the ratio of images diagnosed with tumours by a classifier. The classifier was trained on the GAN generated target dataset, but tested on a balanced holdout test set. It was thereby shown that the generator of CycleGAN effectively learned to hide source domain image features in target domain images, which arguably helped it to fool its discriminator. Paired image-to-image translation with pix2pix (Isola et al., 2017) was more stable, but still some hallucinations were shown to likely have occurred. A cause for this can be a biased discriminator that has learned to discriminating specific image features (e.g., tumours) that are more present in one domain. Cohen et al. (2018a,b) and Wolterink et al. (2018) warn that models that map source to target images, have an incentive to add/remove features during translation if the feature distribution in the target domain is distinct from the feature distribution in the source domain.⁹

Domain-adaptation with unpaired image-to-image translation GANs such as CycleGAN has become increasingly popular in cancer imaging (Wolterink et al., 2017; Tanner et al., 2018; Modanwal et al., 2019; Fossen-Romsaas et al., 2020; Zhao et al., 2020b; Hognon et al., 2019; Mathew et al., 2020; Kearney et al., 2020b; Peng et al., 2020; Jiang et al., 2018; Sandfort et al., 2019). As described, these methods are hallucination-prone and, thus, can put patients at risk when used in clinical settings. More research is needed on how to robustly avoid or detect and eliminate hallucinations in generated data. To this end, we highlight the potential of investigating feature preserving image translation techniques and methods for evaluating whether features have been accurately translated. For instance, in the presence of feature masks or annotations, an additional local reconstruction loss can be introduced in GANs that enforces feature translation in specific image areas.

4.1.5. Data curation and harmonisation

Aside from the limited availability of cancer imaging datasets, a major problem is that the ones available are often not readily useable

and require further curation (Hosny et al., 2018). Curation includes dataset formatting, normalising, structuring, de-identification, quality assessment and other methods to facilitate subsequent data processing steps, one of which is the ingestion of the data into AI models (Diaz et al., 2021). In the past, GANs have been proposed for curation of data labelling, segmentation and annotation of images (details in Section 4.3) and de-identification of facial features, EHRs, etc (details in Section 4.2). Particular to cancer imaging datasets and of significant importance is the correction of artifacts, such as patient motion, metallic objects, chemical shifts and others caused by the image processing pipeline (Pusey et al., 1986; Nehmeh et al., 2002), which run the risk of confusing models with spurious information. Towards the principled removal of artifacts, several GAN solutions have been proposed (Vu et al., 2020b; Koike et al., 2020; Armanious et al., 2020). As for the task of reconstruction of compressed data (e.g., compressed sensing MRI Mardani et al., 2017), markedly, Yang et al. (2018a) proposed DAGAN, which is based on U-Net (Ronneberger et al., 2015), reduces aliasing artifacts, and faithfully preserves texture, boundaries and edges (of brain tumours) in the reconstructed images. Kim et al. (2018a) feed down-sampled high-resolution brain tumour MRI into a GAN framework similar to pix2pix to reconstruct high-resolution images with different contrast. The authors highlight the possible acceleration of MR imagery collection while retaining high-resolution images in multiple contrasts, necessary for further clinical decision-making. As relevant to the context of data quality curation, GANs have also been proposed for image super-resolution in cancer imaging (e.g., for lung nodule detection Gu et al., 2020, abdominal CT You et al., 2019, and breast histopathology Shahidi, 2021).

Beyond the lack of curation, a problem particular to multi-centre studies is that of inconsistent curation between data derived in different centres. These discontinuities arise from different scanners, segmentation protocols, demographics, etc, and can cause significant problems to subsequent ML algorithms that may overfit or bias towards one configuration over another (i.e. acquisition and annotation shifts). GANs have the potential to contribute in this domain as well by bringing the distributions of images across different centres closer together. In this context recent work by Li et al. (2021b) and Wei et al. (2020) used GAN-based volumetric normalisation to reduce the variability of heterogeneous 3D chest CT scans of different slice thickness and dose levels. The authors showed that features in subsequent radiomics analysis exhibit increased alignment. Other works in this

⁹ For example, if one domain contains mainly healthy images, while the other domain contains mainly pathological images.

domain include a framework that could standardise heterogeneous datasets with a single reference image and obtained promising results on an MRI dataset (Hognon et al., 2019), and GANs that learn bidirectional mappings between different vendors to normalise dynamic contrast enhanced (DCE) breast MRI (Modanwal et al., 2019). An interesting research direction to be explored in the future is synthetic multi-centre data generation using GANs, simulating the distribution of various scanners/centres.

4.1.6. Synthetic data assessment

As indicated in Fig. 1(e), a condition of paramount importance is proper evaluation of GAN-generated or GAN-curated data. This evaluation is to verify that synthetic data is useable for a desired downstream task (e.g., segmentation, classification) and/or indistinguishable from real data while ensuring that no private information is leaked. GANs are commonly evaluated based on *fidelity* (realism of generated samples) and *diversity* (variation of generated samples compared to real samples) (Borji, 2021). Different quantitative measures exist to assess GANs based on the fidelity and diversity of its generated synthetic medical images (Yi et al., 2019; Borji, 2021).

Visual Turing tests (otherwise referred to as Visual Assessment, Mean Opinion Score (MOS) Test, and sometimes used interchangeably with In-Silico Clinical Trials) are arguably the most reliable approach, where clinical experts are presented with samples from real and generated data and are tasked to identify which one is generated. Korkinof et al. (2020) showed that their PGGAN-generated (Karras et al., 2017) 1280×1024 mammograms were inseparable by the majority of participants, including trained breast radiologists. A similar visual Turing test was successfully done in the case of skin disease (Ghorbani et al., 2020), super-resolution of CT (You et al., 2019), brain MRI (Kazuhiro et al., 2018; Han et al., 2018), lung cancer CT scans (Chuquicuma et al., 2018), and histopathology images (Levine et al., 2020). For instance, Chuquicuma et al. (2018) trained a DCGAN (Radford et al., 2015) on the LIDC-IDRI dataset (Armato III et al., 2011) to generate 2D (56×56 pixel) pulmonary lung nodule scans that were realistic enough to deceive 2 radiologists with 11 and 4 years of experience. In contrast to computer vision techniques where synthetic data can often be easily evaluated by any non-expert, the requirement of clinical experts makes Visual Turing Tests in this domain much more costly. Furthermore, a lack of scalability and consistency in medical judgement needs to be taken into account as well Brennan and Silman (1992) and visual Turing tests should in the ideal case engage a range of experts to address inter-observer variation in the assessments. Also, iterating over the same observer addresses intra-observer variation—i.e. repeating the process within a certain amount of intervals that could be days or weeks. These problems are further magnified by the shortage of radiology experts (Mahajan and Venugopal, 2020; Rimmer, 2017) which brings up the necessity for supplementary metrics that can automate the evaluation of generative models. Such metrics allow for preliminary evaluation and can enable research to progress without the logistical hurdle of enlisting experts.

Furthermore, in cases where the sole purpose of the generated data is to improve a downstream task—i.e. classification or segmentation—then the prediction success of the downstream task would be the metric of interest. The latter can reasonably be prioritised over other metrics given that the underlying reasons why the synthetic data alters downstream task performance are examined and clarified.¹⁰

¹⁰ For example, synthetic data may balance imbalanced datasets, reduce overfitting on limited training data, or improve model robustness to better capture domain shifts in the test dataset.

Image quality assessment metrics. Wang et al. (2004) have thoroughly investigated image quality assessment metrics. The most commonly applied metrics include structural similarity index measure (SSIM)¹¹ between generated image and reference image (Wang et al., 2004), mean squared error (MSE)¹² and peak signal-to-noise ratio (PSNR).¹³ In a recent example that followed this framework of evaluation, synthetic brain MRI with tumours generated by edge-aware EA-GAN (Yu et al., 2019) was assessed using three such metrics: PSNR, SSIM, and normalised mean squared error (NMSE). The authors integrated an end-to-end sobel edge detector to create edge maps from real/synthetic images that are input into the discriminator in the dEa-GAN variant to enforce improved textural structure and object boundaries. Interestingly, aside from evaluating on the whole image, the authors demonstrated evaluation results focused on the tumour regions, which were overall significantly lower than the whole image. Other works that have evaluated their synthetic images in an automatic manner have focused primarily on the SSIM and PSNR metrics and include generation of CT (Kearney et al., 2020b; Mathew et al., 2020) and PET scans (Zhao et al., 2020b). While indicative of image quality, these similarity-based metrics might not generalise well to human judgement of image similarity, the latter depending on high-order image structure and context (Zhang et al., 2018c). Finding evaluation metrics that are strong correlates of human judgement of perceptual image similarity is a promising line of research. In the context of cancer and medical imaging, we highlight the need for evaluation metrics for synthetic images that correlate with the perceptual image similarity judged by medical experts. Apart from perceptual image similarity, further evaluation metrics in cancer and medical imaging are to be investigated that are able to estimate the diagnostic value of (synthetic) images and, in the presence of reference images, the diagnostic value proportion between target and reference image.

Deep generative model-specific assessment metrics. In recent years, the Inception score (IS) (Salimans et al., 2016) and Fréchet Inception distance (FID) (Heusel et al., 2017) have emerged, offering a more sophisticated alternative for the assessment of synthetic data. The IS uses a classifier to generate a probability distribution of labels given a synthetic image. If the probability distribution is highly skewed, it is indicative that a specific object is present in the image (resulting in a higher IS), while in the case where it is uniform, the image contains a jumble of objects and that is more likely to be non-sense (resulting in a lower IS).¹⁴ The FID metric compares the distance between the synthetic image distribution to that of the real image distribution by comparing extracted high-level features from one of the layers of a classifier (e.g., Inception v3 as in IS). Both metrics have shown promise in the evaluation of GAN-generated data; however, they come with several bias issues that need to be taken into account during evaluation (Chong and Forsyth, 2020; DeVries et al., 2019; Borji, 2019). As these metrics have not been widely used in cancer imaging yet, their applicability on GAN-synthesised cancer images remains to be investigated. In contrast to computer vision datasets containing diverse objects, medical imaging datasets commonly only contain images of one specific organ. In this regard, we promote further research as to how object diversity based methods such as IS can be applied to medical and cancer imaging,

¹¹ SSIM predicts perceived quality and considers image statistics to assess structural information based on luminance, contrast, and structure.

¹² MSE is computed by averaging the squared intensity differences between corresponding pixels of the generated image and the reference image.

¹³ PSNR is an adjustment to the MSE score, commonly used to measure reconstruction quality in lossy compression.

¹⁴ Not only a low label entropy within an image is desired, but also a high label entropy across images: IS also assesses the variety of peaks in the probability distributions generated from the synthetic images, so that a higher variety is indicative of more diverse objects being generated by the GAN (resulting in a higher IS).

which requires, among others, meaningful adjustments of the dataset-specific pretrained classifications models (i.e. Inception v3) that IS and FID rely upon.

Uncertainty quantification as GAN evaluation metric? A general problem facing the adoption of deep learning methods in clinical tasks is their inherent unreliability exemplified by high prediction variation caused by minimal input variation (e.g., one pixel attack [Korpihalkola et al., 2020](#)). This is further exacerbated by the nontransparent decision making process inside deep neural networks thus often described as ‘black box models’ ([Bi et al., 2019](#)). Also, the performance of deep learning methods in out-of-domain datasets has been assessed as unreliable ([Lim et al., 2019](#)). To eventually achieve beneficial clinical adoption and trust, examining and reporting the inherent uncertainty of these models on each prediction becomes a necessity. Besides classification, segmentation ([Hu et al., 2020](#); [Alshehhi and Alshehhi, 2021](#)), etc, uncertainty estimation is applicable to models in the context of data generation as well ([Lim et al., 2019](#); [Abdar et al., 2020](#); [Hu et al., 2020](#)). [Edupuganti et al. \(2019\)](#) studied a GAN architecture based on variational autoencoders (VAE) ([Kingma and Welling, 2013](#)) on the task of MRI reconstruction, with emphasis on uncertainty studies. Due to their probabilistic nature, VAEs allowed for a Monte Carlo sampling approach which enables quantification of pixel-variance and the generation of uncertainty maps. Furthermore, they used Stein’s Unbiased Risk Estimator (SURE) ([Stein, 1981](#)) as a measure of uncertainty that serves as surrogate of MSE even in the absence of ground truth. Their results indicated that adversarial losses introduce more uncertainty. Parallel to image reconstruction, uncertainty has also been studied in the context of brain tumours (glioma) in MRI enhancement ([Tanno et al., 2021](#)). In this study, a probabilistic deep learning framework for model uncertainty quantification was proposed, decomposing the problem into two uncertainty types: *intrinsic uncertainty* (particular to image enhancement and pertaining to the one-to-many nature of the super-resolution mapping) and *parameter uncertainty* (a general challenge, it pertains to the choice of the optimal model parameters). The overall model uncertainty in this case is a combination of the two and was evaluated for image super-resolution. Through a series of systematic studies the utility of this approach was highlighted, as it resulted in improved overall prediction performance of the evaluated models even for out-of-distribution data. It was further shown that predictive uncertainty highly correlated with reconstruction error, which not only enabled spotting unrealistic synthetic images, but also highlights the potential in further exploring uncertainty as an evaluation metric for GAN-generated data. A further use-case of interest for GAN evaluation via uncertainty estimation is the ‘adherence’ to provided conditional inputs. As elaborated in 4.1.4 for image-to-image translation, conditional GANs are likely to introduce features that do not correspond to the conditional class label or source image. After training a classification model on image features of interest (say, tumour vs non-tumour features), we can examine the classifier’s prediction and estimated uncertainty¹⁵ for the generated images. Given the expected features in the generated images are known beforehand, the classifier’s uncertainty of the presence of these features can be used to estimate not only image fidelity (e.g., image features are not generated realistic enough), but also ‘conditional adherence’ (e.g., expected image features are altered during generation).

Outlook on clinical adoption. Alongside GAN-specific and standard image assessment metrics, uncertainty-based evaluation schemes can further automate the analysis of generative models. To this end, the challenge of clinical validation for predictive uncertainty as a reliability metric for synthetic data assessment remains ([Tanno et al., 2021](#)). In

practice, building clinical trust in AI models is a non-trivial endeavour and will require rigorous performance monitoring and calibration especially in the early stages ([Kelly et al., 2019](#); [Durán and Jongsma, 2021](#)). This is particularly the case when CAde and CADx models are trained on entirely (or partially) synthetic data given that the data itself was not first assessed by clinicians. Until a certain level of trust is built in these pipelines, automatic metrics will be a preliminary evaluation step that is inevitably followed by diligent clinical evaluation for deployment. A research direction of interest in this context would be ‘gatekeeper’ GANs—i.e. GANs that simulate common data (and/or difficult edge cases) of the target hospital, on which deployment-ready candidate models (e.g., segmentation, classification, etc.) are then tested to ensure they are sufficiently generalisable. If the candidate model performance on such test data satisfies a predefined threshold, it has passed this quality gate for clinical deployment.

4.2. Data access and privacy challenges

Access to sufficiently large and labelled data resources is the main constraint for the development of deep learning models for medical imaging tasks ([Esteve et al., 2019](#)). In cancer imaging, the practice of sharing validated data to aid the development of AI algorithms is restricted due to technical, ethical, and legal concerns ([Bi et al., 2019](#)). The latter is exemplified by regulations such as the Health Insurance Portability and Accountability Act (HIPAA, 1996) in the United States of America (USA) or the European Union’s General Data Protection Regulation (GDPR, 2016) with which respective clinical centres must comply with. Alongside the need and numerous benefits of patient privacy preservation, it can also limit data sharing initiatives and restrict the availability, size and usability of public cancer imaging datasets. [Bi et al. \(2019\)](#) assess the absence of such datasets as a noteworthy challenge for AI in cancer imaging.

The published GANs and adversarial training methods that are suggested for or applied to cancer imaging challenges within this Section 4.2 are summarised below in Table 3.

4.2.1. Decentralised data generation

As AI systems are often developed and trained outside of medical institutions, prior approval to transfer data out of their respective data silos is required, adding significant hurdles to the logistics of setting up a training pipeline or rendering it entirely impossible. In addition, medical institutions can often not guarantee a secured connection to systems deployed outside their centres ([Hosny et al., 2018](#)), which further limits their options to share valuable training data.

One privacy preserving approach solving this problem is federated learning ([McMahan et al., 2017](#)), where copies of an AI model are trained in a distributed fashion inside each clinical centre in parallel and are aggregated to a global model in a central server. This eliminates the need for sensitive patient data to leave any of the clinical centres ([Kaissis et al., 2020](#); [Sheller et al., 2020](#)). However, it is to be noted that federated learning cannot guarantee full patient privacy. [Hitaj et al. \(2017\)](#) demonstrated that any malicious user can train a GAN to violate the privacy of the other users in a federated learning system. While difficult to avoid, the risk of such GAN-based attacks can be minimised, e.g., by using a combination of selective parameter updates ([Shokri and Shmatikov, 2015](#)) (sharing only a small selected part of the model parameters across centres) and the sparse vector technique¹⁶ as shown by [Li et al. \(2019b\)](#).

To solve the challenge of privacy assurance of clinical medical imaging data, a distributed GAN ([Hardy et al., 2019](#); [Xin et al., 2020](#); [Guerraoui et al., 2020](#); [Rasouli et al., 2020](#); [Zhang et al., 2021](#)) can

¹⁵ The uncertainty can be estimated using methods such as Bayesian Neural Networks ([MacKay, 1992](#); [Neal, 2012](#)), Monte-Carlo Dropout ([Gal and Ghahramani, 2016](#)) or Deep Ensembles ([Lakshminarayanan et al., 2016](#)).

¹⁶ Sparse Vector Technique (SVT) ([Lyu et al., 2016](#)) is a Differential Privacy (DP) ([Dwork, 2006](#)) method that introduces noise into a deep learning model’s gradients.

Table 2

Overview of adversarially-trained models applied to cancer imaging **data scarcity** and **usability** challenges. Publications are clustered by section and ordered by year in ascending order.

Publication	Method	Dataset	Modality	Task	Highlights
Imbalanced data & fairness					
Hu et al. (2018b)	ProstateGAN	Private	Prostate MRI	Class-conditional synthesis	Gleason score (cancer grade) class conditions.
Ghorbani et al. (2020)	DermGAN	Private	Dermoscopy	Paired translation	Adapted pix2pix evaluated via Turing Tests and rare skin condition CLF.
Li et al. (2021a)	Encoder	ISIC 2018 (Codella et al., 2018)	Dermoscopy	Adversarial training, Representation learning	Fair Encoder with bias discriminator and skin lesion CLF.
Cross-modal data generation					
Wolterink et al. (2017)	CycleGAN	Private	Cranial MRI/CT	Unpaired translation	First CNN for unpaired MR-to-CT translation. Evaluated via PSNR and MAE.
Ben-Cohen et al. (2017)	pix2pix	Private	Liver PET/CT	Paired translation	Paired CT-to-PET translation with focus on hepatic malignant tumours.
Nie et al. (2017)	context-aware GAN	ADNI (Wyman et al., 2013; Weiner et al., 2017)	Cranial/pelvic MRI/CT	Paired translation	Supervised 3D GAN for MR-to-CT translation with 'Auto-Context Model' (ACM).
Wang et al. (2018b)	Locality Adaptive GAN (LA-GAN)	BrainWeb phantom (Cocosco et al., 1997)	Cranial MRI, PET phantom	Paired translation	3D auto-context, synthesising PET from low-dose PET and multimodal MRI.
Tanner et al. (2018)	CycleGAN	VISCERAL (Jimenez-del Toro et al., 2016)	Lung/abdominal MRI/CT	Image registration	MR-CT CycleGAN for registration.
Kaiser and Albarqouni (2019)	pix2pix, context-aware GAN (Nie et al., 2017)	RIRE (Fitzpatrick, 1998)	Cranial MRI/CT	Paired translation	Detailed preprocessing description, MR-to-CT translation, comparison with U-Net.
Prokopenko et al. (2019)	DualGAN, SRGAN	CPTAC3 (National Cancer Institute, 2018) & Head-Neck-PET-CT (Valli��res et al., 2017)	Cranial MRI/CT	Unpaired translation	DualGAN for unpaired MR-to-CT, visual Turing tests.
Yurt et al. (2019) mrrecon	mustGAN	IXI (IXI Dataset, 2007) & ISLES (Maier et al., 2017)	Cranial MRI	Paired translation	FLAIR, T1, T2 synthesis via feature fusion of one-to-one and many-to-one pix2pix networks.
Li et al. (2019a)	diamondGAN	Private & MICCAI-WMH (Kuijff et al., 2019)	Cranial MRI	Unpaired translation	Target modality synthesis from flexible set of non-aligned source modalities.
Zhou et al. (2020)	hi-Net	BRATS 2018 (Menze et al., 2014; Bakas et al., 2018)	Cranial MRI	Paired translation	Domain-specific encoder network features fused into layers of pix2pix-based network.
Zhao et al. (2020b)	S-CycleGAN	Private	Cranial low/full dose PET	Paired translation	Low (LDPET) to full dose (FDPET) translation with supervised loss for paired images.
Kearney et al. (2020b)	VAE-enhanced A-CycleGAN	Private	Cranial MRI/CT	Unpaired translation	MR-to-CT evaluated via PSNR, SSIM, MAE. Superior to paired alternatives.
Kazemifar et al. (2020)	context-aware GAN	Private	Cranial MRI/CT	Paired translation	Feasibility of generated CT from MRI for dose calculation for radiation treatment.
Armanious et al. (2020)	MedGAN	Private	Liver PET/CT	Paired translation	CasNet architecture, PET-to-CT, MRI motion artifact correction, PET denoising.
Xu et al. (2020)	multi-conditional GAN	NSCLC (Zhou et al., 2018)	Lung CT, gene expression	Multi-input conditional synthesis	Image-gene data fusion, nodule generator input: background, segmentation, gene code.
Haubold et al. (2021)	Pix2PixHD (Wang et al., 2018a)	Private	Arterial phase CT	Paired translation	Low-to-full ICM CT (thorax, liver, abdomen), 50% reduction in intravenous ICM dose.

(continued on next page)

Table 2 (continued).

Publication	Method	Dataset	Modality	Task	Highlights
Feature hallucinations					
Cohen et al. (2018a,b) dist-bias	CycleGAN, pix2pix	BRATS2013 (Menze et al., 2014)	Cranial MRI	Paired/unpaired translation	Removed/added tumours during image translation can lead to misdiagnosis.
Data curation					
Yang et al. (2018a)	DAGAN	MICCAI 2013 grand challenge dataset	Cranial MRI	Image reconstruction	Fast GAN compressed sensing MRI reconstruction outperformed conventional methods.
Kim et al. (2018a)	pix2pix-based	BRATS (Menze et al., 2014)	Cranial MRI	Reconstruction/super-resolution	Information transfer between different contrast MRI, effective pretraining/fine-tuning.
Hognon et al. (2019)	CycleGAN, pix2pix	BRATS (Menze et al., 2014), BrainWeb phantom (Cocosco et al., 1997)	Cranial MRI	Paired/unpaired translation, normalisation	CycleGAN translation to BrainWeb reference image, pix2pix back-translation to source.
Modanwal et al. (2019)	CycleGAN	Private	Breast MRI	Unpaired translation	Standardising DCE-MRI across scanners, anatomy preserving mutual information loss.
You et al. (2019)	CycleGAN-based	Mayo Low Dose CT (AAPM, 2016)	Abdominal CT	Super-resolution	Joint constraints to Wasserstein loss for structural preservation. Evaluated by 3 radiologists.
Gu et al. (2020)	MedSRGAN	LUNA16 (Setio et al., 2017)	MRI/thoracic CT	Super-resolution	Residual Whole Map Attention Network (RWMAN) in G. Evaluated by 5 radiologists.
Vu et al. (2020b)	WGAN-GP-based	k-Wave toolbox (Treeby and Cox, 2010)	Photoacoustic CT (PACT)	Paired translation	U-Net & WGAN-GP based generator for artifact removal. Evaluated via SSIM, PSNR.
Koike et al. (2020)	CycleGAN	Private	Head/neck CT	Unpaired translation	Metal artifact reduction via CT-to-CT translation, evaluated via radiotherapy dose accuracy.
Wei et al. (2020)	WGAN-GP-inspired	Private	Chest CT	Paired translation	CT normalisation of dose/slice thickness. Evaluated via Radiomics Feature Variability.
Shahidi (2021)	WA-SRGAN	BreakHis (Benhammou et al., 2020), Camelyon (Litjens et al., 2018)	Breast/lymph node histopathology	Super-resolution	Wide residual blocks, self-attention SRGAN for improved robustness & resolution
Li et al. (2021b)	SingleGAN-based (Yu et al., 2018b)	Private	Spleen/colorectal CT	Unpaired translation	Multi-centre (4) CT normalisation. Evaluated via cross-centre radiomics features variation. Short/long-term survivor CLF improvement.
Synthetic data assessment					
Kazuhiro et al. (2018)	DCGAN	Private	Cranial MRI	Noise-to-image synthesis	Feasibility study for brain MRI synthesis evaluated by 7 radiologists.
Han et al. (2018)	DCGAN, WGAN	BRATS 2016 (Menze et al., 2014)	Cranial MRI	Noise-to-image synthesis	128 × 128 brain MRI synthesis evaluated by one expert physician.
Chuquicusma et al. (2018)	DCGAN	LIDC-IDRI (Armato III et al., 2011)	Thoracic CT	Noise-to-image synthesis	Malignant/benign lung nodule ROI generation evaluated by two radiologists.
Yu et al. (2019)	Ea-GAN	BRATS 2015 (Menze et al., 2014), IXI (IXI Dataset, 2007)	Cranial MRI	Paired image-to-image translation	Loss based on edge maps of synthetic images. Evaluated via PSNR, NMSE, SSIM.
Korkinof et al. (2020)	PGGAN	Private	Full-field digital MMG	Noise-to-image synthesis	1280 × 1024 MMG synthesis from > 10 ⁶ image dataset. Evaluated by 55 radiologists.
Levine et al. (2020)	PGGAN, VAE, ESRGAN	TCGA (Grossman et al., 2016), OVCARE archive	Ovarian Histopathology	Noise-to-image synthesis	1024 × 1024 whole slide synthesis. Evaluated via FID and by 15 pathologists (9 certified)

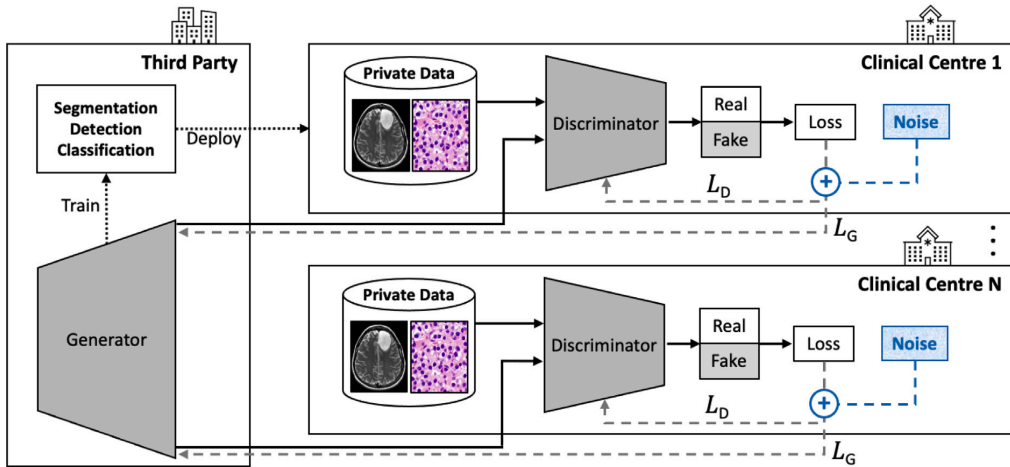


Fig. 7. Visual example of a GAN in a federated learning setup with a central generator trying to generate realistic samples that fool all of the discriminators, which are distributed across clinical centres as in Chang et al. (2020b,a). Once trained, the generator can produce training data for a downstream task model (e.g., segmentation, detection, classification). As depicted in blue colour, we suggest to extend the federated learning setup by adding 'Noise' to the gradients providing a differential privacy guarantee. This reduces the possibility of reconstruction of specific records of the training data (i.e. images of a specific patient) by someone with access to the trained GAN model (i.e. to the generator) or by someone intercepting the synthetic images while they are transferred from the central generator to the centres during training.

be trained on sensitive patient data to generate synthetic training data. The technical, legal, and ethical constraints for sharing de-identified synthetic data are typically less restrictive than for real patient data. Such generated data can be used instead of the real patient data to train models on disease detection, segmentation, or prognosis.

For instance, Chang et al. (2020b,a) proposed the Distributed Asynchronous Discriminator GAN (AsynDGAN), which consists of multiple discriminators deployed inside various medical centres and one central generator deployed outside the medical centres. The generator never needs to see the private patient data, as it learns by receiving the gradient updates of each of the discriminators. The discriminators are trained to differentiate images of their medical centre from synthetic images received from the central generator. After training AsynDGAN, its generator is used and evaluated based on its ability to provide a rich training set of images to successfully train a segmentation model. AsynDGAN is evaluated on MRI brain tumour segmentation and cell nuclei segmentation. The segmentation models trained only on AsynDGAN-generated data achieves a competitive performance when compared to segmentation models trained on the entire dataset of real data. Notably, models trained on AsynDGAN-generated data outperform models trained on local data from only one of the medical centres. To our best knowledge, AsynDGAN is the only distributed GAN applied to cancer imaging to date. Therefore, we promote further research in this line to fully exploit the potential of privacy-preservation using distributed GANs. As demonstrated in Fig. 7 and suggested in Fig. 1(f), for maximal privacy preservation we recommend exploring methods that combine privacy during training (e.g., federated GANs) with privacy after training (e.g., differentially-private GANs), the latter being described in the following section.

4.2.2. Differentially-private data generation

Shin et al. (2018a) train a GAN to generate brain tumour images and highlight the usefulness of their method for anonymisation, as their synthetic data cannot be attributed to a single patient but rather only to an instantiation of the training population. However, it is to be scrutinised whether such synthetic samples are indeed fully private, as, given a careful analysis of the GAN model and/or its generated samples, a risk of possible reconstruction of part of the GAN training data exists (Papernot et al., 2016). For example, Chen et al. (2020a) propose a GAN for model inversion (MI) attacks, which aim at reconstructing

the training data from a target model's parameters. A potential solution to avoid training data reconstruction is highlighted by Xie et al. (2018), who propose the Differentially Private Generative Adversarial Network (DPGAN). In Differential Privacy (DP) (Dwork, 2006) the parameters (ϵ, δ) denote the privacy budget (Torfi et al., 2020), where ϵ measures the privacy loss and δ represents the probability that a range of outputs with a privacy loss $> \epsilon$ exists.¹⁷ Hence, the smaller the parameters (ϵ, δ) for a given model, the less effect a single sample in the training data has on model output. The less effect of such a single sample, the stronger is the confidence in the privacy of the model to not reveal samples of the training data.

Examples of GANs with differential privacy guarantees. In DPGAN noise is added to the model's gradients during training to ensure training data privacy. Extending on the concept of DPGAN, Jordon et al. (2018) train a GAN coined PATE-GAN based on the Private Aggregation of Teacher Ensembles (PATE) framework (Papernot et al., 2016, 2018). In the PATE framework, a student model learns from various unpublished teacher models each trained on data subsets. The student model cannot access an individual teacher model nor its training data. PATE-GAN consists of k discriminator teachers, $T_1, \dots, T_k$, and a student discriminator S that backpropagates its loss back into the generator. This limits the effect of any individual sample in PATE-GAN's training. In a $(\epsilon = 1, \delta = 10^{-5})$ -DP setting, classification models trained on PATE-GAN's synthetic data achieves competitive performances e.g. on a *non-imaging* cervical cancer dataset (Fernandes et al., 2017) compared to an upper bound vanilla GAN baseline without DP.

On the same dataset, Torfi et al. (2020) achieve competitive results using a Rényi Differential Privacy and Convolutional Generative Adversarial Networks (RDP-CGAN) under an equally strong $(\epsilon = 1, \delta = 10^{-5})$ -DP setting.

For the generation of biomedical participant data in clinical trials, Beaulieu-Jones et al. (2019) apply an AC-GAN under a $(\epsilon = 3.5, \delta =$

¹⁷ For example, if an identical model M is trained two times, once with training data D resulting in M_D and once with marginally different training data D' resulting in $M_{D'}$, it is (ϵ) -DP if the following holds true: For any possible output x , the output probability for x of model M_D differs no more than $\exp(\epsilon)$ from the output probability for x of $M_{D'}$.

10^{-5})-DP setting based on Gaussian noise added to AC-GAN's gradients during training.

Bae et al. (2020) propose AnomiGAN to anonymise private medical data via some degree of output randomness during inference. This randomness of the generator is achieved by randomly adding, for each layer, one of its separately stored training variances. AnomiGAN achieves competitive results on a *non-imaging* breast cancer dataset and a *non-imaging* prostate cancer for any of the reported privacy parameter values $\epsilon \in [0.0, 0.5]$ compared to DP, where Laplacian noise is added to samples.

Outlook on synthetic cancer image privacy. Despite the above efforts, DP in GANs has only been applied to non-imaging cancer data which indicates research potential for extending these methods reliably to cancer imaging data. According to Stadler et al. (2021), using synthetic data generated under DP can protect outliers in the original data from linkage attacks, but likely also reduces the statistical signal of these original data points, which can result in lower utility of the synthetic data. Apart from this privacy-utility tradeoff, it may not be readily controllable/predictable which original data features are preserved and which omitted in the synthetic datasets (Stadler et al., 2021). In fields such as cancer imaging where patient privacy is critical, desirable privacy-utility tradeoffs need to be defined and thoroughly evaluated to enable trust, shareability, and usefulness of synthetic data. Consensus is yet to be found as to how privacy preservation in GAN-generated data can be evaluated and verified in the research community and in clinical practice. Promising approaches include methods that define a privacy gain/loss for synthetic samples (Stadler et al., 2021; Yoon et al., 2020). Yoon et al. (2020), for instance, define and backpropagate an identifiability loss to the generator to synthesis anonymised electronic health records (EHRs). The identifiability loss is based on the notion that the minimum weighted euclidean distance between two patient records from two different patients can serve as a desirable anonymisation target for synthetic data. Designing or extending reliable methods and metrics for standardised quantitative evaluation of patient privacy preservation in synthetic medical images is a line of research that we call attention to.

4.2.3. Obfuscation of identifying patient features in images

If the removal of all sensitive patient information within a cancer imaging dataset allows for sharing such datasets, then GANs can be used to obfuscate such sensitive data. As indicated by Fig. 1(g), GANs can learn to remove the features from the imaging data that could reveal a patient's identity, e.g. by learning to apply image inpainting to pixel or voxel data of burned in image annotations or of identifying body parts. Such identifying body parts could be the facial features of a patient, as was shown by Schwarz et al. (2019) on the example of cranial MRI. Numerous studies exist where GANs accomplish facial feature de-identification on non-medical imaging modalities (Wu et al., 2018b; Hukkelås et al., 2019; Li and Lin, 2019; Maximov et al., 2020). For medical imaging modalities, GANs have yet to prove themselves as tool of choice for anatomical and facial feature de-identification against common standards (Ségonne et al., 2004; Bischoff-Grethe et al., 2007; Schimke et al., 2011; Milchenko and Marcus, 2013) with solid baselines. These standards, however, have shown to be susceptible to reconstruction achieved by unpaired image-to-image GANs on MRI volumes with high reversibility for blurred faces and partial reversibility for removed facial features (Abramian and Eklund, 2019). Van der Goten et al. (2021) provide a first proof-of-concept for GAN-based facial feature de-identification in 3D (128^3 voxel) cranial MRI. The generator of their conditional de-identification GAN (C-DeID-GAN) receives brain mask, brain intensities and a convex hull of the brain MRI as input and generates de-identified MRI slices. C-DeID-GAN generates the entire de-identified brain MRI scan and, hence, may not be able to guarantee that the generation process does not alter any of the original brain features. A solution to this can be to only generate and replace the 2D MRI slices

or parts thereof that do contain non-pathological facial features while retaining all other original 2D MRI slices. Presuming preservation of medically relevant features and robustness of de-identification, GAN-based approaches can allow for subsequent medical analysis, privacy preserving data sharing and provision of de-identified training data. Hence, we highlight the research potential of GANs for robust medical image de-identification e.g. via image inpainting GANs that have already been successful applied to other tasks in cancer imaging such as synthetic lesion inpainting into mammograms (Wu et al., 2018a; Becker et al., 2019) and lung CT scans (Mirsky et al., 2019). Also, GAN-based patient feature de-identification methods that are adjustable and trainable to remain quantifiably robust against adversarial image reconstruction are a research line of interest.

4.2.4. Identifying patient features in latent representations

In line with Fig. 1(g), a further example for privacy preserving methods are autoencoders¹⁸ that learn patient identity-specific features and obfuscate such features when encoding input images into latent space representation. Such an identity-obfuscated representation can be used as input into further models (classification, segmentation, etc.) or decoded back into a de-identified image. Adversarial training has been shown to be effective for learning a privacy-preserving encoding function, where a discriminator tries to succeed at classifying the private attribute from the encoded data (Raval et al., 2017; Wu et al., 2018c; Yang et al., 2018c; Pittaluga et al., 2019). Apart from being trained via the backpropagated adversarial loss, the encoder needs at least one further utility training objective to learn to generate useful representations, such as denoising (Vincent et al., 2008) or classification of a second attribute (e.g., facial expressions Chen et al., 2018a; Oleszkiewicz et al., 2018). Siamese Neural Networks (Bromley et al., 1994) such as the Siamese Generative Adversarial Privatizer (SGAP) (Oleszkiewicz et al., 2018) have been used effectively for adversarial training of an identity-obfuscated representation encoder. In SGAP, two weight-sharing Siamese Discriminators are trained using a distance based loss function to learn to classify whether a pair of images belongs to the same person. As visualised in Fig. 8, Kim et al. (2019a) follow a similar approach with the goal of de-identifying and segmenting brain MRI data. Two feature maps are encoded from a pair of MRI scans and fed into a Siamese Discriminator that evaluates via binary classification whether the two feature maps are from the same patient. The generated feature maps are also fed into a segmentation model that backpropagates a Dice loss (Sudre et al., 2017) to train the encoder. Fig. 8 illustrates the scenario where the encoder is deployed in a trusted setting after training, e.g. in a clinical centre, and the segmentation model is deployed in an untrusted setting, e.g. outside the clinical centre at a third party. The encoder shares the identity-obfuscated feature maps with the external segmentation model without the need of transferring the sensitive patient data outside the clinical centre. This motivates further research into adversarial identity-obfuscated encoding methods e.g., to allow sharing and usage of cancer imaging data representations and models across clinical centres.

4.2.5. Adversarial attacks putting patients at risk

Examples of GAN-based tampering with cancer imaging data. For instance, Mirsky et al. (2019) added and removed evidence of cancer in lung CT scans. Of two identical deep 3D convolutional cGANs (based on pix2pix), one was used to inject (diameter ≥ 10 mm) and the other to remove (diameter < 3 mm) multiple solitary pulmonary nodules indicating lung cancer. The GANs were trained on 888 CT scans from the Lung Image Database Consortium image collection (LIDC-IDRI) dataset (Armato III et al., 2011) and inpainted on an extracted

¹⁸ For example adversarial autoencoders (Makhzani et al., 2015; Creswell et al., 2018), which adversarially learn latent space representations that match a chosen prior distribution.

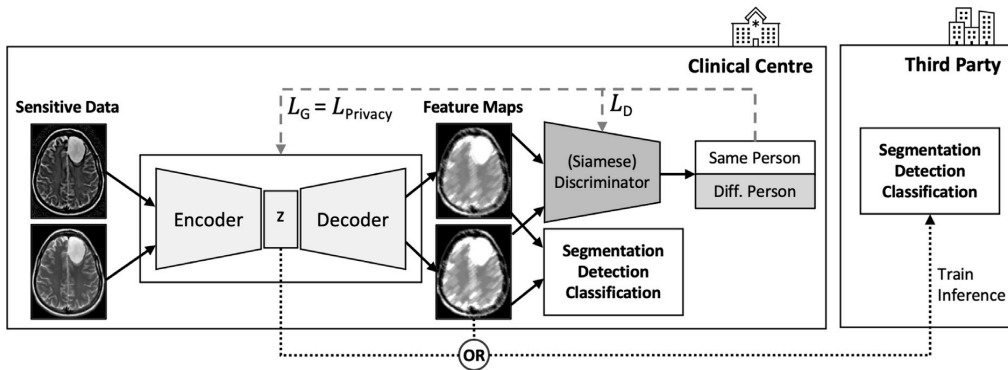


Fig. 8. Example of an autoencoder architecture trained via adversarial loss to learn privacy-preserving feature maps as in Kim et al. (2019a) and/or a privacy-preserving latent representation z . Once trained and after thorough periodic manual verification of its ability to preserve privacy, the representation z and/or the feature maps can be sent to third parties outside the clinical centre for model training or inference requests.

region of interest of 32^3 voxel cuboid shape. The trained GANs can be autonomously executed by malware and are capable of ingesting nodules into standard CT scans that are realistic enough to deceive both radiologists and AI disease detection systems. Three radiologists with 2, 5 and 7 years of experience analysed 70 tampered and 30 authentic CT scans. Spending on average 10 min on each scan, the radiologists diagnosed 99% of the scans with added nodules as malignant and 94% of the scans with removed nodules as healthy. After disclosing the presence of the attack to the radiologists, the percentages dropped to 60% and 87%, respectively (Mirsky et al., 2019).

Becker et al. (2019) trained a CycleGAN (Zhu et al., 2017) on 680 down-scaled mammograms from the Breast Cancer Digital Repository (BCDR) (Lopez et al., 2012) and the INbreast (Moreira et al., 2012) datasets to generate suspicious features and was able to remove or inject them into existing mammograms. They showed that their approach can fool radiologists at lower pixel dimensions (i.e. 256×256) demonstrating that alterations in patient images by a malicious attacker can remain undetected by clinicians, influence the diagnosis, and potentially harm the patient (Becker et al., 2019).

Defending adversarial attacks. In regard to fooling diagnostic models, one measure to circumvent adversarial attacks is to increase model robustness against adversarial examples (Madry et al., 2017), as suggested by Fig. 1(h). Augmenting the robustness has been shown to be effective for medical imaging segmentation models (He et al., 2019; Park et al., 2020), lung nodule detection models (Liu et al., 2020b; Paul et al., 2020), skin cancer recognition (Huq and Pervin, 2020; Hirano et al., 2021), and classification of histopathology images of lymph node sections with metastatic tissue (Wetstein et al., 2020). Liu et al. (2020b) provide model robustness by adding adversarial chest CT examples to the training data. These adversarial examples are composed of synthetic nodules that are generated by a 3D convolutional variational encoder trained in conjunction with a WGAN-GP (Gulrajani et al., 2017) discriminator. To further enhance robustness, Projected Gradient Descent (PGD) (Madry et al., 2017) is applied to find and protect against noise patterns for which the detector network is prone to produce over-confident false predictions (Liu et al., 2020b).

Apart from being the adversary, GANs can also detect adversarial attacks and thus are applicable as security counter-measure enabling attack anticipation, early warning, monitoring and mitigation. DefenceGAN, for example, learns the distribution of non-tampered images and can generate a close output to an inference input image that does not contain adversarial modifications (Samangouei et al., 2018).

We highlight the research potential in adversarial attacks and examples, alongside prospective GAN detection and defence mechanisms that can, as elaborated, highly impact the field of cancer imaging.

Apart from the image injection of entire tumours and the generation of adversarial radiomics examples, a further attack vector to consider in future studies is the perturbation of the specific imaging features within an image that are used to compute radiomics features.

4.3. Data annotation and segmentation challenges

4.3.1. Annotation-specific issues in cancer imaging

Missing annotations in datasets. In cancer imaging, not only the availability of large datasets is rare, but also the availability of labels, annotations, and segmentation masks within such datasets. The generation and evaluation of such labels, annotations, and segmentation masks is a task for which trained health professionals (radiologists, pathologists) are needed to ensure validity and credibility (Hosny et al., 2018; Bi et al., 2019). Nonetheless, radiologist annotations of large datasets can take years to generate (Bi et al., 2019). The tasks of labelling and annotating (e.g., bounding boxes, segmentation masks, textual comments) cancer imaging data is, hence, expensive both in time and cost, especially considering the large amount of data needed to train deep learning models.

Intra/inter-observer annotation variability. This cancer imaging challenge is further exacerbated by the high intra- and inter-observer variability between both pathologists (Gilles et al., 2008; Dimitriou et al., 2018; Martin et al., 2018; Klaver et al., 2020) and radiologists (Elmore et al., 1994; Hopper et al., 1996; Hadjiiski et al., 2012; Teh et al., 2017; Wilson et al., 2018; Woo et al., 2020; Brady, 2017) in interpreting cancer images across imaging modalities, affected organs, and cancer types. Automated annotation processes based on deep learning models allow to produce reproducible and standardised results in each image analysis. In one of most common case where the annotations consist of a segmentation mask, reliably segmenting both tumour and non-tumour tissues is crucial for disease analysis, biopsy, and subsequent intervention and treatment (Hosny et al., 2018; Huynh et al., 2020), the latter being further discussed in Section 4.5. For example, automatic tumour segmentation models are useful in the context of radiotherapy treatment planning (Cuocolo et al., 2020).

Human biases in cancer image annotation. During routine tasks, such as medical image analysis, humans are prone to account for only a few of many relevant qualitative image features. On the contrary, the strength of GANs and deep learning models is the evaluation of large numbers of multi-dimensional image features alongside their (non-linear) inter-relationships and combined importance (Hosny et al., 2018). Deep learning models are likely to react to unexpected and subtle patterns in the imaging data (e.g., anomalies, hidden comorbidities, etc.) that

Table 3

Overview of adversarially-trained models applied/applicable to **data access and privacy** cancer imaging challenges. Publications are clustered by section and ordered by year in ascending order.

Publication	Method	Dataset	Modality	Task	Highlights
Decentralised GANs					
Chang et al. (2020b,a) AsynDGAN	AsynDGAN, PatchGAN (Isola et al., 2017)	BRATS 2018 (Bakas et al., 2018), Multi-Organ (Kumar et al., 2017)	Cranial MRI, nuclei images	Paired translation	Mask-to-image, central G gets distributed Ds' gradients, synthetic only-trained segmentation.
Differential-privacy GANs					
Xie et al. (2018)	DPGAN	MNIST (LeCun et al., 1998), MIMIC-III (Johnson et al., 2016)	MNIST images, EHRs	Noise-to-image synthesis	Noisy gradients during training ensure DP guarantee.
Jordon et al. (2018)	PATE-GAN	Cervical cancer (Fernandes et al., 2017)	[non-imaging] Medical records	Data synthesis	DP via PATE framework. G gradient from student D that learns from teacher Ds.
Beaulieu-Jones et al. (2019)	AC-GAN	MIMIC-III (Johnson et al., 2016)	[non-imaging] EHRs, clinical trial data	Class-conditional synthesis	DP via Gaussian noise added to AC-GAN gradient. Treatment arm (standard/intensive) as condition.
Bae et al. (2020)	AnomiGAN	UCI b & Prostate (Blake, 1998)	[non-imaging] Cell nuclei tabular data	Multi-class-conditional synthesis, classification	DP via training variances added to G's layers in inference. Real data row as G's condition.
Torfi et al. (2020)	RDP-CGAN	Cervical cancer (Fernandes et al., 2017), MIMIC-III (Johnson et al., 2016)	[non-imaging] Medical records, EHRs	Data synthesis	DP GAN based on Rényi divergence. Allows to track a DP loss.
Patient de-identification					
Abramian and Eklund (2019)	CycleGAN	IXI (IXI Dataset, 2007)	Cranial MRI	Unpaired translation	Reconstruction of blurring/removed faces in MRI shows privacy vulnerability.
Kim et al. (2019a)	PrivacyNet	PPMI (Marek et al., 2011)	Cranial MRI	Adversarial training, segmentation	Segmenting de-identified representations learned via same-person CLF by Siamese Ds.
Van der Goten et al. (2021)	C-DeID-GAN	ADNI (Wyman et al., 2013; Weiner et al., 2017), OASIS-3 (LaMontagne et al., 2019)	Cranial MRI	Paired translation	Face de-id. Concatenated convex hull, brain mask & brain volumes as G & D inputs.
Adversarial data tampering					
Mirsky et al. (2019)	pix2pix-based CT-GAN	LIDC-IDRI (Armato III et al., 2011)	Lung CT	Image inpainting	Injected/removed lung nodules in CT fool radiologists and AI models.
Becker et al. (2019)	CycleGAN	BCDR (Lopez et al., 2012), INbreast (Moreira et al., 2012)	Digital/Film MMG	Unpaired image-to-image translation	Suspicious features can be learned and injected/removed from MMG.
Liu et al. (2020b)	Variational Encoder, WGAN-GP	LUNA (Setio et al., 2017), NLST (NLST Research Team, 2011)	Lung CT	Noise-to-image synthesis	Robustness via adversarial data augmentation, reduce false positives in nodule detection.

medical practitioners are prone to overlook for instance due to any of multiple existing cognitive biases (e.g., anchoring bias, framing bias, availability bias) (Brady, 2017) or inattentional blindness (Drew et al., 2013). Inattentional blindness occurs when radiologists (or pathologist) have a substantial amount of their attention drawn to a specific task, such as finding an expected pattern (e.g., a lung nodule) in the imaging data, that they become blind to other patterns in that data.

Implications of low segmentation model robustness. As for the common annotation task of segmentation mask delineation, automated segmentation models can minimise the risk of the aforesaid human biases. However, to date, segmentation models have difficulties when confronted with intricate segmentation problems including domain shifts, rare diseases with limited sample size, or small lesion and

metastasis segmentation. In this sense, the performance of many automated and semi-automated clinical segmentation models has been sub-optimal (Sharma and Aggarwal, 2010). This emphasises the need for expensive manual verification of segmentation model results by human experts (Hosny et al., 2018). The challenge of training automated models for difficult segmentation problems can be approached by applying methods for learning discriminative features without explicit labels. Such methods include GANs and variational autoencoders (Kingma and Welling, 2013) capable of automating robust segmentation (Hosny et al., 2018).

In addition, segmented regions of interest (ROI) are commonly used to extract quantitative imaging features with diagnostic value such as radiomics features. The latter are used to detect and monitor tumours

Table 4

Overview of adversarial training and GAN-based approaches applied to **segmentation** in cancer imaging tasks. Publications are clustered by organ type and ordered by year in ascending order. “*” indicates that the metrics are only available in figures and the baseline numbers are lower than using GANs in the corresponding paper. ‘n.a.’ indicates that there was no comparison with a specific baseline with the reason for this being indicated in the ‘Highlights’ column.

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+Δ)	Highlights
Head/Brain/Neck							
Kamnitsas et al. (2017) deepmedic	deepmedic	Private	MRI	Adversarial training	Dice: 0.60 Recall: 0.56 Precision: 0.70	0.63, 0.59, 0.72	Multi-domain (MPRAGE, FLAIR, T2, Proton Density, and Gradient-Echo) segmentation with domain discriminator loss.
Rezaei et al. (2017)	pix2pix, MarkovianGAN (Li and Wand, 2016)	BRATS 2017 (Menze et al., 2014; Bakas et al., 2018, 2017)	MRI	Adversarial training	Dice: n.a.	0.80	D detects G-generated masks for high/low grade glioma segmentation. Benchmarked in a challenge, thus missing metric w/o GAN.
Mok and Chung (2018)	CB-GAN	BRATS (Menze et al., 2014)	MRI	Data augmentation	Dice: 0.79	0.84	Coarse-to-fine G captures training set manifold, generates generic samples in HGG & LGG segmentation.
Yu et al. (2018a)	pix2pix-based	BRATS (Menze et al., 2014)	MRI	Data augmentation	Dice: 0.67	0.68	Cross-modal paired FLAIR to T1 translation, training tumour segmentation with T1+real/synthetic FLAIR. Baseline: only-T1 training
Shin et al. (2018a)	pix2pix	BRATS (Menze et al., 2014)	MRI	Data augmentation	Dice: 0.81	0.81	Training on synthetic, before fine-tuning on 10% of the real data.
Xue et al. (2018)	SegAN	BRATS (Menze et al., 2014)	MRI	Adversarial training	Dice: 0.80	0.85	Paired image-to-mask. New multi-scale loss: L1 of D representations between GT- & prediction-masked input MRI. U-Net baseline.
Giacomello et al. (2020)	SegAN-CAT	BRATS (Menze et al., 2014)	MRI	Adversarial training	Dice: 0.71	0.86	Paired image-to-mask. Combined dice loss & multi-scale loss using concatenation on channel axis instead of masking. SegAN baseline.
Kim et al. (2020) BrainTumor	GAN	BRATS (Bakas et al., 2018; Menze et al., 2014)	MRI	Image inpainting, data augmentation	Dice: 0.57	0.59	Simplifying tumour features into concentric circle & grade mask to inpaint.
Hu et al. (2020)	UNet-based GAN segmenter	Private	CT, PET	Adversarial training	Dice: 0.69	0.71	Spatial context information & hierarchical features. Nasal-type lymphoma segmentation with uncertainty estimation.
Qasim et al. (2020)	SPADE-based	BRATS (Bakas et al., 2018), ISIC (Codella et al., 2019)	MRI, dermoscopy	Cross-domain translation	Dice: *	B:0.66 S:0.62	Brain and skin segmentation. Frozen segmenter as 3rd player in GAN to condition on local apart from global information.
Foroozandeh and Eklund (2020)	PGGAN, SPADE	BRATS (Menze et al., 2014)	MRI	Data augmentation	Av. dice error(%): 16.80	16.18	Sequential noise-to-mask and paired mask-to-image translation to synthesise labelled tumour images.
Cirillo et al. (2020)	vox2vox: 3D pix2pix	BRATS (Menze et al., 2014)	MRI	Adversarial training	Dice: 0.87	0.93	3D adversarial training to enforce segmentation results to look realistic.
Han et al. (2021)	Symmetric adaptation network	BRATS (Menze et al., 2014)	MRI	Cross-domain translation	Dice: 0.48	0.67	Simultaneous source/target (T2 to other sequences) translation and segmentation. Compared to CycleGAN baseline.
Alshehhi and Alshehhi (2021)	U-Net based GAN segmenter	BRATS (Menze et al., 2014)	MRI	Adversarial training	Dice: n.a.	n.a.	Quantitative comparison of 7 active learning acquisition functions using existing adversarial networks (including SegAN, SegAN-CAT)
Breast							
Singh et al. (2018)	pix2pix	DDSM (Heath et al., 2001)	Film MMG	Adversarial training	Dice: 0.86	0.94	Adversarial loss to make automated segmentation close to manual masks for breast mass segmentation.
Caballo et al. (2020)	DCGAN (Radford et al., 2015)	Private	CT	Data augmentation	Dice: 0.70	0.93	GAN based data augmentation; Validated by matching extracted radiomics features.
Negi et al. (2020)	GAN, WGAN-RDA-UNET	Private	Ultrasound	Adversarial training	Dice: 0.85	0.88	Outperforms state-of-the-art using Residual-Dilated-Attention-Gate-UNet and WGAN for lesion segmentation.
Abdominal							
Huo et al. (2018)	Conditional PatchGAN	Private	MRI	Adversarial training	Dice: 0.93	0.94	Adversarial loss as segmentation post-processing for spleen segmentation. ResNet baseline.
Chen et al. (2019)	DC-FCN-based GAN segmenter	LiTS (Bilic et al., 2019)	CT	Adversarial training	Dice: 0.62	0.68	Cascaded framework with densely connected adversarial training. DC-Fully connected network baseline.

(continued on next page)

Table 4 (continued).

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+ Δ)	Highlights
Sandfort et al. (2019)	CycleGAN	NIH (Prior et al., 2017), Decathlon (Simpson et al., 2019), DeepLesion (Yan et al., 2018)	CT	Cross-domain translation	Dice (od): 0.92, 0.10	0.93, 0.75	Contrast enhanced to non-enhanced translation to improve out-of-distribution (od) segmentation.
Xiao et al. (2019)	Radiomics-guided GAN	Private	MRI	Adversarial training	Dice: 0.72	0.92	Radiomics-guided adversarial mechanism to map relationship between contrast and non-contrast images. U-Net baseline comparison.
Oliveira (2020b)	pix2pix, SPADE	LiTS (Bilic et al., 2019)	CT	Image inpainting	Dice: 0.58	0.61	Realistic lesion inpainting in CT slices to provide controllable set of training samples.
Chest and lungs							
Jiang et al. (2018)	CycleGAN-based	NSCLC (Prior et al., 2017)	CT, MRI	Cross-domain translation	Dice: 0.66	0.80	Tumour-aware loss for unsupervised cross-domain adaptation compared with standard cycleGAN benchmark.
Jin et al. (2018)	cGAN	LIDC (Armato III et al., 2011)	CT	Image inpainting	Dice: 0.96	0.99	Generated lung nodules to improve segmentation robustness; A novel multi-mask reconstruction loss.
Dong et al. (2019)	UNet-based GAN segmenter	AAPM Challenge (Yang et al., 2018b)	CT	Adversarial training	Dice: 0.97 (l), 0.83 (sc), 0.71 (o), 0.85 (h)	0.97, 0.90, 0.75, 0.87	Adversarial training to discriminate manual and automated segmentation of lungs, spinal cord, oesophagus, heart.
Tang et al. (2019) XLSor	MUNIT (Huang et al., 2018)	JSRT (Shiraishi et al., 2000), Montgomery (Jaeger et al., 2013)	Chest X-ray	Lung segmentation	Dice: 0.97	0.98	Unpaired normal-to-abnormal (pathological) translation. Synthetic data augmentation for lung segmentation.
Shi et al. (2020)	AUGAN	LIDC-IDRI (Armato III et al., 2011)	CT	Adversarial training	Dice: 0.82	0.85	A deep layer aggregation based on U-Net++.
Munawar et al. (2020)	Unet-based GAN segmenter	JSRT (Shiraishi et al., 2000), Montgomery & Shenzhen (Jaeger et al., 2013)	Chest X-ray	Adversarial training	Dice: 0.96	0.97	Adversarial training to discriminate manual and automated segmentation.
Prostate							
Kohl et al. (2017)	UNet-based GAN segmenter	Private	MRI	Adversarial training	Dice: 0.35	0.41	Adversarial loss to discriminate manual and automated segmentation.
Grall et al. (2019) ProstatecGAN	pix2pix	Private	MRI	Adversarial training	Dice: 0.67 (ADC), 0.74 (DWI), 0.67 (T2)	0.73, 0.79, 0.74	Paired prostate image-to-mask translation. Investigated the robustness of the pix2pix against noise.
Nie and Shen (2020)	GAN	PROMISE12 (Litjens et al., 2014)	MRI	Adversarial confidence learning	Dice: 88.25 (b), 90.11 (m), 86.67 (a)	89.52, 90.97, 88.20	Difficulty-aware mechanism to alleviate the effect of easy samples during training. b = base, m = middle, a = apex.
Zhang et al. (2020)	PGGAN	Private	CT	Data augmentation	Dice: 0.85	0.90	Semi-supervised training using both annotated and un-annotated data. Synthetic data augmentation using PGGAN.
Cem Birbiri et al. (2020)	pix2pix, CycleGAN	PROMISE12 (Litjens et al., 2014)	MRI	Data augmentation	Dice: 0.72	0.76	Compared the performance of pix2pix, U-Net, and CycleGAN.
Chaitanya et al. (2021)	cGAN, DCGAN-based D	ACDC (Bernard et al., 2018), Decathlon (Simpson et al., 2019)	MRI, CT	Data augmentation	Dice: 0.40	0.53	GAN data augmentation for intensity and shape variations. Tested on cardiac, prostate, and pancreas datasets.
Colorectal							
Liu et al. (2019a)	GAN, LAGAN	Private	CT	Adversarial training	Dice: 0.87	0.92	Automatic post-processing to refine the segmentation of deep networks.
Poorneshwaran et al. (2019)	pix2pix	CVC-clinic (Vázquez et al., 2017)	Endoscopy	Adversarial training	Dice: n.a.	0.88	Adversarial learning to make automatic segmentation close to manual. Ablations were compared instead of baselines.
Xie et al. (2020)	MI ² GAN, CycleGAN	CVC-clinic (Vázquez et al., 2017), ETIS-Larib (Silva et al., 2014)	Endoscopy	Cross-domain translation	Dice: 0.66	0.73	Content features and domain information decoupling and maximising the mutual information.

(continued on next page)

Table 4 (continued).

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+4)	Highlights
Pathology							
Pandey et al. (2020)	GAN & cGAN	Kaggle (Ljosa et al., 2012a)	Histopathology	Data augmentation	Dice: 0.79	0.83	Two-stage GAN to generate masks and conditioned synthetic images.
Other							
Chi et al. (2018)	pix2pix	ISBI ISIC (Codella et al., 2018)	Dermoscopy	Data augmentation	Dice: 0.85	0.84	Similar performance replacing half of real with synthetic data. Colour labels as lesion specific characteristics.
Abhishek and Hamarneh (2019)	pix2pix	ISBI ISIC (Codella et al., 2018)	Dermoscopy	Data augmentation	Dice: 0.77	0.81	Generate new lesion images given any arbitrary mask.
Sarker et al. (2019)	MobileGAN	ISBI ISIC (Codella et al., 2018)	Dermoscopy	Adversarial training	Dice: 0.76	0.91	Lightweight and efficient GAN model with position and channel attention. Dice was compared to a U-Net baseline.
Zaman et al. (2020)	pix2pix	Private	Ultrasound	Data augmentation	Dice: *	≈ 4+0.05	Recommendations on standard data augmentation approaches. Bone surface segmentation. pix2pix based discriminator.

(e.g., lymphoma Kang et al., 2018), biomarkers, and tumour-specific phenotypic attributes (Lambin et al., 2012; Parmar et al., 2015). The accuracy and success of such commonly applied diagnostic image feature quantification methods, hence, depends on accurate and robust ROI segmentations. Segmentation models need to be able to provide reproducibility of extracted quantitative features and biomarkers (Bi et al., 2019) with reliably-low variation, among others, across different scanners, CT slice thicknesses, and reconstruction kernels (Balagurunathan et al., 2014; Zhao et al., 2016). To this end, we promote lines of research that use adversarial training schemes to target the robustification of segmentation models. Progress in this open research challenge can beneficially unlock trust, usability, and clinical adoption of biomarker quantification methods in clinical practice.

4.3.2. GAN applications for cancer image segmentation

Table 4 summarises the collection of segmentation publications that utilise such adversarial training approaches and GAN-based data synthesis for cancer imaging.

In Table 4, we further report the baseline performance alongside the performance increase attributable to applying GANs or adversarial training for each surveyed publication. For the common Dice Score segmentation performance metric, Fig. 10 visualises these differences. Comparing the figure’s black identity line and the red trend line over publications, we observe a general improvement of approximately 5 percentage points of adversarial learning methods compared to their baselines. Fig. 11 further displays the variation in performance between baselines and adversarial network methods for the years 2017 to 2021. Based on visual analysis, performance gains seem to be both anatomy-invariant and invariant to the strength of the baseline, where similar gains are achieved for initially low (e.g., < 0.7) and high (e.g., ≥ 0.7) baseline Dice scores. While Figs. 10 and 11 offer interesting quantitative insights, we recommend taking potential publication bias¹⁹ into account when drawing conclusions from these figures. Trends in the presented data in these plots can be analysed holistically, however, they are not intended to benchmark and compare individual publications against each other. This is due to multiple limiting factors of such comparisons including the differences in (a) the used baselines, (b) organs, (c) modalities, (d) the segmentation task and its associated difficulty, (e) the amount of training and testing data, (f) data and annotation quality, (g) pre- and post-processing methods, or (h) the study’s objectives. In regard to (g), some studies may focus on other

benefits of adversarial learning methods instead of or apart from Dice Score improvement, such as, reducing the needed training dataset size, domain adaptation in general, protecting patient privacy with synthetic data, or simply improving other metrics (e.g. Hausdorff distance, FID).

In the following sections, we provide a summary of the commonly used techniques and trends in the GAN literature that address the challenges in cancer image segmentation.

Robust quantitative imaging feature extraction. For example, Xiao et al. (2019) addressed the challenge of robustification of segmentation models and reliable biomarker quantification. Xiao et al. (2019) provide radiomics features as conditional input to the discriminator of their adversarially trained liver tumour segmentation model. Their learning procedure strives to inform the generator to create segmentations that are specifically suitable for subsequent radiomics feature computation. Apart from adversarially training segmentation models, we also highlight the research potential of adversarially training quantitative imaging feature extraction models (e.g., deep learning radiomics) for reliable application in multi-centre and multi-domain settings.

Synthetic segmentation model training data. By augmenting and varying the training data of segmentation models, it is possible to substantially decrease the amount of manually annotated images during training while maintaining the performance (Foroozandeh and Eklund, 2020). A general pipeline of such usage of GAN based generative models is demonstrated in Fig. 9(a) and mentioned in Fig. 1(j).

Over the past few years, CycleGAN (Zhu et al., 2017) based approaches have been widely used for synthetic data generation due to the possibility of using unpaired image sets in training, as compared to paired image translation methods like pix2pix (Isola et al., 2017) or SPADE (Park et al., 2019). CycleGAN based data augmentation has been shown to be useful for segmentation model training, in particular, for generating images with different acquisition characteristics such as contrast enhanced MRI from non-contrast MRI (Wang et al., 2021), cross-modality image translation between different modalities such as CT and MRI images (Huo et al., 2018), and domain adaptation tasks (Jiang et al., 2018). The popularity of the CycleGAN based methods lies not only in image synthesis or domain adaptation, but also in the inclusion of simultaneous image segmentation in its pipeline (Lee et al., 2020).

Although pix2pix methods require paired samples, it is also a widely used type of GAN in data augmentation for medical image segmentation (see Table 4). Several works on segmentation have demonstrated its effectiveness in generating synthetic medical images. By manipulating its input, the variability of the training dataset for image segmentation could be remarkably increased in a controlled manner (Abhishek and

¹⁹ Publication bias likely influences the trends observable in Figs. 10 and 11: Only papers that show an improvement attributable to adversarial networks were published and therefore only such studies could be included.

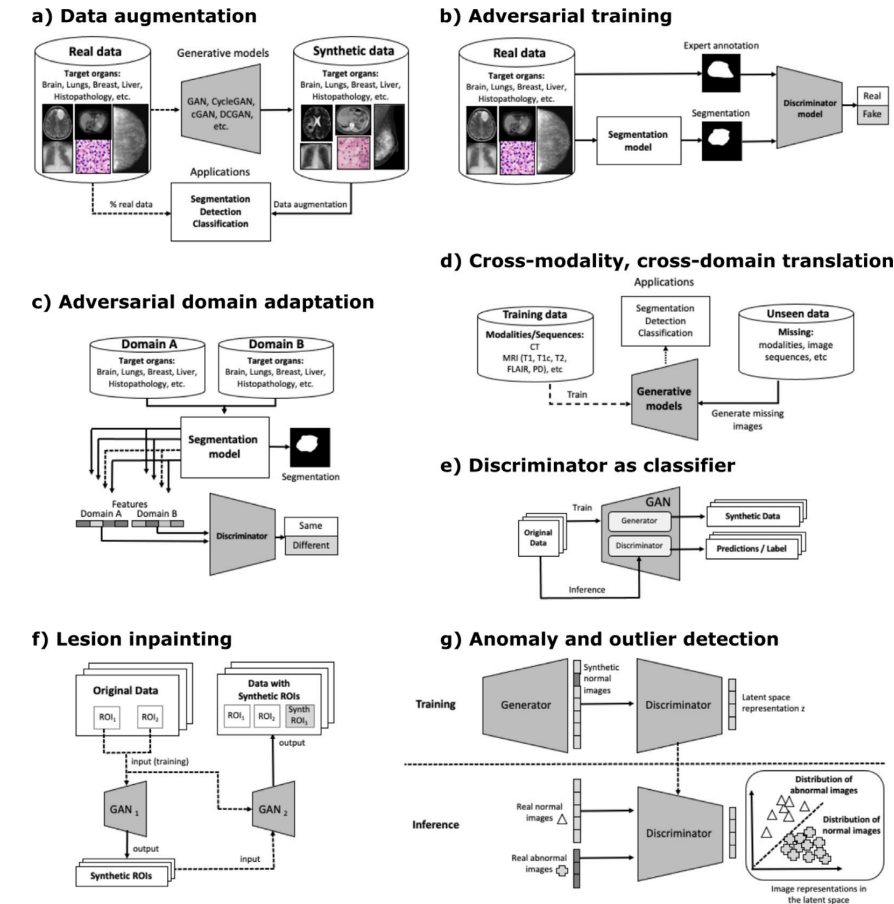


Fig. 9. Overview of cancer imaging GAN applications for detection and segmentation. (a) describes training data augmentation of downstream task models (e.g., segmentation, detection, classification, etc.). In (b) a discriminator scrutinises the segmentations created by a segmentation model, while in (c) the discriminator enforces the model to create domain-agnostic latent representations. (d) illustrates target domain adaptation, where the translated target domain images are used for downstream model training. In (e), the AC-GAN (Odena et al., 2017) discriminator classifies original data. In (f), one GAN generates ROIs while another inpaints them into full-sized images. (g) uses the discriminator's latent space to find abnormal/outlier representations.

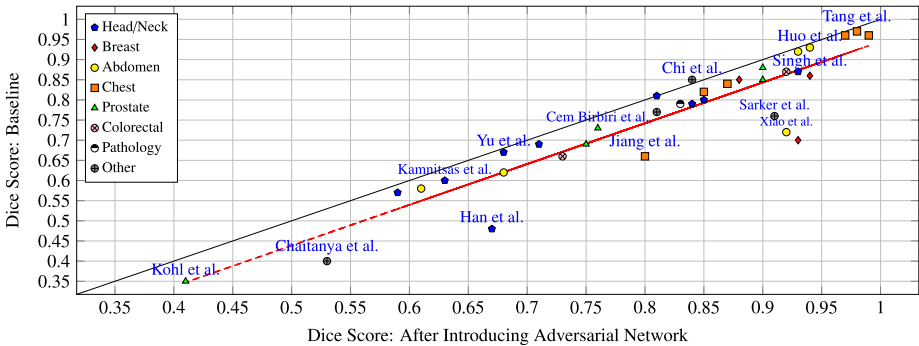


Fig. 10. Scatter plot illustrating the segmentation performance improvement attributable to adversarial networks for the surveyed publications. Each publication is represented by a marker with a colour and shape encoding depicting the publication's anatomical category. Only the publications are included that measure performance via Dice Score and compare against a baseline, as reported in Table 4. For publications reporting multiple Dice Scores, their mean was computed and included herein. The black identity line indicates no change between baseline and adversarial network intervention, while dots below this line represent an improvement. The red regression line depicts the trend of improvement across publications. The author names of a few publications have been manually selected for highlighting based on the distance to the trend line.

Table 5

Overview of adversarially-trained models applied to **detection and diagnosis** tasks in cancer imaging. Publications are clustered by organ type and ordered by year in ascending order.

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+Δ)	Highlights
Brain							
Chen et al. (2018b)	VAE GAN	Cam-CAN (Shafro et al., 2014) BRATS (Menze et al., 2014) ATLAS (Liew et al., 2018)	MRI	Anomaly detection	AUC(%): 80.0	70.0	Comparison of unsupervised outlier detection methods.
Han et al. (2020)	PGGAN	BRATS (Menze et al., 2014)	MRI	Data augmentation	Accuracy(%): 90.06	91.08	PGGAN-based augmentation method of whole brain MRI.
Han et al. (2019c)	PGGAN, SimGAN	BRATS (Menze et al., 2014)	MRI	Data augmentation	Sensitivity(%): 93.67	97.48	Two-step GAN for noise-to-image and image-to-image data augmentation.
Benson and Beets-Tan (2020)	GAN	TCIA (Schmainda and Prah, 2018)	MRI	Anomaly detection	Accuracy(%): 73.48	74.96	Multi-modal MRI images as input to the GAN.
Han et al. (2019b)	GAN	Private	MRI	Data augmentation	Sensitivity(%): 67.0	77.0	Synthesis and detection of brain metastases in MRI with bounding boxes.
Siddiquee et al. (2019)	Fixed-Point GAN	BRATS (Menze et al., 2014)	MRI	Anomaly detection	AUC(%): 95, IoU:0.261	92, 0.348	Brain lesion detection and localisation with fixed-point image-to-image translation concept.
Sun et al. (2020b)	ANT-GAN	BRATS 2013 (Menze et al., 2014)	MRI	Data augmentation, Anomaly detection	F1-Score(%): 89.6	91.7	Abnormal to normal image translation in cranial MRI, VGG lesion classification.
Breast							
Wu et al. (2018a) Mammo-ciGAN	ciGAN	DDSM (Heath et al., 2001)	Film MMG	Data augmentation	ROC AUC(%): 88.7	89.6	Synthetic lesion in-filling in healthy mammograms.
Guan and Loew (2019)	GAN	DDSM (Heath et al., 2001)	Film MMG	Data augmentation	Accuracy(%): 73.48	74.96	Generate patches containing benign and malignant tumours.
Jendele et al. (2019) BreastGAN	CycleGAN	BCDR (Lopez et al., 2012), INbreast (Moreira et al., 2012), CBIS-DDSM (Lee et al., 2016)	Digital/Film MMG	Data augmentation	ROC AUC(%): 83.50, F1(%): 62.53	4–1.46, 4+1.28	Scanned & digital mammograms evaluated together for lesion detection.
Lee and Nishikawa (2020)	CGAN	Private	Digital MMG	Data augmentation	ROC AUC(%): 57	67	Synthesising contralateral mammograms.
Wu et al. (2020)	U-Net based GAN	OPTIMAM (Halling-Brown et al., 2020)	Digital MMG	Data augmentation	ROC AUC(%): 82.9	84.6	Removed/added MMG lesions for malignant/benign classification.
Alyafi et al. (2020)	DCGAN	OPTIMAM (Halling-Brown et al., 2020)	Digital MMG	Data augmentation	F1-Score(%): ≈90	≈96	Synthesise breast mass patches with high diversity.
Desai et al. (2020)	DCGAN	DDSM (Heath et al., 2001)	Film MMG	Data augmentation	Accuracy(%): 78.23	87.0	Synthesise full view MMGs and used them in visual Turing test.
Muramatsu et al. (2020)	CycleGAN	DDSM (Heath et al., 2001)	CT, Film MMG	Data augmentation	Accuracy(%): 79.2	81.4	CT lung nodule to MMG mass translation and vice versa.
Swiderski et al. (2021)	AGAN	DDSM (Heath et al., 2001)	Film MMG	Data augmentation	ROC AUC(%): 92.50	94.10	AutoencoderGAN (AGAN) augments data in normal abnormal classification.
Kansal et al. (2020)	DCGAN	Private	OCT	Data augmentation	Accuracy(%): 92.0	93.7	Synthetic Optical Coherence Tomography (OCT) images.
Shen et al. (2021)	ciGAN	DDSM (Heath et al., 2001)	Film MMG	Data augmentation	FROC-AUC(%): 15.1	17.2	Generate labelled breast mass images for precise detection.
Pang et al. (2021)	TripleGAN-based	Private	Ultrasound	Data augmentation	Sensitivity(%): 86.60	87.94	Semi-supervised GAN-based Radiomics model for mass CLF.
Liver							
Frid-Adar et al. (2018)	DCGAN, ACGAN	Private	CT	Data augmentation	Sensitivity(%): 78.6	85.7	Synthesis of high quality focal liver lesions from CT for lesion CLF.
Zhao et al. (2020a)	Tripartite GAN	Private	MRI	Data augmentation	Accuracy(%): 79.2	89.4	Synthetic contrast-enhanced MRI → tumour detection without contrast agents.

(continued on next page)

Table 5 (continued).

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+Δ)	Highlights
Doman et al. (2020)	DCGAN	JAMIT (JAMIT Japanese Society of Medical Imaging Technology), 3Dircadb (Soler et al., 2010)	CT	Data augmentation	Detection rate(%): 65	95	Generate metastatic liver lesions in abdominal CT for improved cancer detection.
Stomach/Colon/Prostate							
Kanayama et al. (2019)	DCGAN	Private	Endoscopy	Data augmentation	AP(%): 59.6	63.2	Synthesise lesion images for gastric cancer detection.
Shin et al. (2018b)	cGAN	CVC-CLINIC (Bernal et al., 2015), CVC-ClinicVideoDB (Angermann et al., 2017)	Colonoscopy	Data augmentation	Precision(%): 81.9	85.3	Synthesise polyp images from normal colonoscopy images for polyp detection.
Rau et al. (2019) ColonoscopyDepth	pix2pix-based	Private	Colonoscopy	Data augmentation	Mean L1(%): 33.9	24.7	Transform monocular endoscopic images from two domains to depth maps.
Yu and Zhang (2020)	CapGAN	BrainWeb phantom (Cocosco et al., 1997), Prostate MRI (Choyke et al., 2016)	MRI	Data augmentation	ROC AUC(%): 85.1	88.5	Synthesise prostate MRI using Capsule Network-Based DCGAN instead of CNN.
Krause et al. (2021)	CGAN	TCGA, NLCS (van den Brandt et al., 1990)	Histopathology	Data augmentation	ROC AUC(%): 75.7	77.7	GANs to enhance genetic alteration detection in colorectal cancer histology.
Skin							
Bissoto et al. (2018) gan-skin-lesion	PGAN, DCGAN, pix2pix	Dermofit (Ballerini et al., 2013), ISIC 2017 (Codella et al., 2018), IAD (Argenziano et al., 2002)	Dermoscopy	Data augmentation	ROC AUC(%): 83.4	84.7	Comparative study of GANs for skin lesions generation
Creswell et al. (2018)	ssDAE	ISIC 2017 (Codella et al., 2018)	Dermoscopy	Representation learning, classification	ROC AUC(%): 89.0	89.0	Adversarial autoencoder fine-tuned on few labelled lesion classification samples.
Baur et al. (2018)	DCGAN, LAPGAN	ISIC 2017 (Codella et al., 2018)	Dermoscopy	Data augmentation	Accuracy(%) : 71.6	74.0	Comparative study, 256 × 256 px skin lesions synthesis. LAPGAN acc = 74.0%
Rashid et al. (2019)	GAN	ISIC 2017 (Codella et al., 2018)	Dermoscopy	Data augmentation	Accuracy(%): 81.5	86.1	Boost CLF performance with GAN-based skin lesions augmentation.
Fossen-Romsaas et al. (2020)	AC-GAN, CycleGAN	HAM10000 & BCN20000 (Tschandl et al., 2018; Combalia et al., 2019), ISIC 2017 (Codella et al., 2018)	Dermoscopy	Data augmentation	Recall(%): 72.1	76.3	Realistic-looking, class-specific synthetic skin lesions.
Qin et al. (2020)	Style-based GAN	ISIC 2017 (Codella et al., 2018)	Dermoscopy	Data augmentation	Precision(%): 71.8	76.9	Style control & noise input tuning in G to synthesise high quality lesions for CLF.
Lung							
Bi et al. (2017)	M-GAN	Private	PET	Data augmentation	F1-Score(%): 66.38	63.84	Synthesise PET data via multi-channel GAN for tumour detection.
Salehinejad et al. (2018)	DCGAN	Private	Chest X-rays	Data augmentation	Accuracy(%): 70.87	92.10	Chest pathology CLF using synthetic data.
Zhao et al. (2018)	F(&)BGAN	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	Accuracy(%): 92.86	95.24	Forward GAN generates diverse images, Backward GAN improves their quality.
Onishi et al. (2019)	WGAN	Private	CT	Data augmentation	Accuracy(%): 63 (Benign), 82 (Malign)	67, 94	Synthesise pulmonary nodules on CT images for nodule CLF.

(continued on next page)

Table 5 (continued).

Publication	Method	Dataset	Modality	Task	Metric w/o GAN (Baseline)	Metric with GAN (Baseline+Δ)	Highlights
Gao et al. (2019) 3DGANLungNodules	WGAN-GP	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	Sensitivity(%): 84.8	95.0	Synthesise lung nodule 3D data for nodule detection.
Han et al. (2019a)	3DMCGAN	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	CPM(%): 51.8	55.0	3D multi-conditional GAN (2 Ds) for misdiagnosis prevention in nodule detection.
Yang et al. (2019)	GAN	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	ROC AUC(%): 87.56	88.12	Class-aware 3D lung nodule synthesis for nodule CLF.
Wang et al. (2020b) CA-MW-AS	CGAN	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	F1-Score(%): 85.88	89.03	Nodule synthesis conditioned on semantic features.
Kuang et al. (2020)	Multi-D GAN	LIDC-IDRI (Armato III et al., 2011)	CT	Anomaly detection	Accuracy(%): 91.6	95.32	High anomaly scores on malignant images, low scores on benign.
Ghosal et al. (2020)	WGAN-GP	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	ROC AUC(%): 95.0	97.0	Unsupervised AE & clustering augmented learning method for nodule CLF.
Sun et al. (2020a)	DCGAN	LIDC-IDRI (Armato III et al., 2011)	CT	Data augmentation	Accuracy(%): 93.8	94.5	Nodule CLF: Pre-training AlexNet (Krizhevsky et al., 2012) on synthetic, fine-tuning on real.
Wang et al. (2020c)	pix2pix, PGWGAN, WGAN-GP	Private	CT	Data augmentation	Accuracy(%): 53.2	60.5	Augmented CNN for subcentimeter pulmonary adenocarcinoma CLF.
Bu et al. (2020)	3D CGAN	LUNA16 (Setio et al., 2017)	CT	Data augmentation	Sensitivity(%): 97.81	98.57	Squeeze-and-excitation mechanism and residual learning for nodule detection.
Nishio et al. (2020)	3D pix2pix	LUNA16 (Setio et al., 2017)	CT	Data augmentation	Accuracy(%): 85	85	Nodule size CLF. Masked image + mask + nodule size conditioned paired translation.
Onishi et al. (2020)	WGAN	Private	CT	Data augmentation	Specificity(%): 66.7	77.8	AlexNet pretrained on synthetic, fine-tuned on real nodules for malign/benign CLF.
Teramoto et al. (2020)	PGGAN	Private	Cytopathology	Data augmentation	Accuracy(%): 81.0	85.3	Malignancy CLF: CNN pretrained on synthetic lung cytology images, fine-tuned on real.
Others							
Schlegl et al. (2017)	AnoGAN	Private	OCT	Anomaly detection	ROC AUC(%): 73	89	D representations trained on healthy retinal image patches to score abnormal patches.
Zhang et al. (2018b)	DCGANs, WGAN, BEGANs	Private	OCT	Data augmentation	Accuracy(%): 95.67	98.83	CLF of thyroid/non-thyroid tissue. Comparative study for GAN data augmentation.
Chaudhari et al. (2019)	MG-GAN	NCBI (Edgar et al., 2002)	Expression microarray data	Data augmentation	Accuracy(%) P:71.43 L:68.21 B:69.8 C:67.59	93.6, 88.1, 90.3, 91.7	Prostate, Lung, Breast, Colon. Interesting for fusion with imaging data.
Liu et al. (2019b)	WGAN-based	Private	Serum sample staging data	Data augmentation	Accuracy(%): 64.52	70.97	Synthetic training data for CLF of stages of Hepatocellular carcinoma.
Rubin et al. (2019)	TOP-GAN	Private	Holographic microscopy	Data augmentation	AUC(%): 89.2	94.7	Pretrained D adapted to CLF of optical path delay maps of cancer cells (colon, skin).

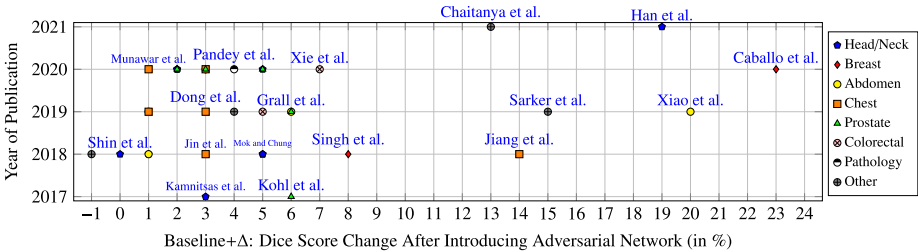


Fig. 11. Scatter plot displaying year of publication and Dice Score improvement (in %). Each marker represents a publication and its colour and shape encoding represents its corresponding anatomical category. Only the publications are included that report Dice Score alongside a baseline comparison. For publications reporting multiple Dice Scores, their mean was computed and included herein. Author names have been manually selected at random for highlighting.

Hamarneh, 2019; Oliveira, 2020b). Similarly, the conditional GAN methods have also been used for controllable data augmentation for improving lesion segmentation (Oliveira, 2020a). Providing a condition

as an input to generate a mask is particularly useful to specify the location, size, shape, and heterogeneity of the synthetic lesions. One of the recent examples, proposed by Kim et al. (2020), demonstrates

this in brain MRI tumour synthesis by conditioning an input with simplified controllable concentric circles to specify lesion location and characteristics. A further method for data augmentation is the inpainting of generated lesions into healthy real images or into other synthetic images, as depicted by Fig. 9(f). Overall, the described data augmentation techniques have shown to improve generalisability and performance of segmentation models by increasing both the number and the variability of training samples (Qasim et al., 2020; Foroozandeh and Eklund, 2020; Lee et al., 2020).

Segmentation models with integrated adversarial loss. As stated in Fig. 1(i), GANs can also be used as the algorithm that generates robust segmentation masks, where the generator is used as a segmenter and the discriminator scrutinises the segmentation masks given an input image. One intuition behind this approach is the detection and correction of higher-order inconsistencies between the ground truth segmentation maps and the ones created by the segmenter via adversarial learning (Luc et al., 2016; Hu et al., 2020; Cirillo et al., 2020). This approach is demonstrated in Fig. 9(b). With the additional adversarial loss when training a segmentation model, this approach has been shown to improve semantic segmentation accuracy (Hung et al., 2019; Sarker et al., 2019; Shi et al., 2020). Using adversarial training, similarity of a generated mask to manual segmentation given an input image is taken under consideration by the discriminator allowing a global assessment of the segmentation quality. This approach further offers a practical solution towards handling intra- and inter-observer annotation variability, as the mask discriminator learns an average over observers, which is backpropagated to the segmenter via adversarial loss.

A unique way of incorporating the adversarial loss from the discriminator has been recently proposed in Nie and Shen (2020). In their work, the authors utilise a fully-convolutional network as a discriminator, unlike its counterparts that use binary, single neuron output networks. In doing so, a dense confidence map is produced by the discriminator, which is further used to train the segmenter with an attention mechanism.

Overall, using an adversarial loss as an additional global segmentation assessment is likely to be a helpful further signal for segmentation models, in particular, for heterogeneously structured datasets of limited size (Kohl et al., 2017), as is common for cancer imaging datasets. We highlight potential further research in GAN-based segmentation models to learn to segment increasingly fine radiologic distinctions. These models can help to solve further cancer imaging challenges, for example, accurate differentiation between neoplasms and tissue response to injury in the regions surrounding a tumour after treatment (Bi et al., 2019).

Segmentation models with integrated adversarial domain discrimination. Moreover, a similar adversarial loss can also be performed internally on the segmentation model features as illustrated in Fig. 9(c). Such an approach can benefit unsupervised domain adaptation and domain generalisation by enforcing the segmentation model to learn to base its prediction on domain-invariant feature representations (Kamnitsas et al., 2017).

4.3.3. Limitations and future prospects for cancer imaging segmentation

As shown in Table 4, the applications of GANs in cancer image segmentation cover a variety of clinical requirements. Remarkable steps have been taken to advance this field of research over the past few years. However, the following limitations and future prospects can be considered for further investigation:

- Although the data augmentation using GANs could increase the number of training samples for segmentation, the variability of the synthetic data is limited to the training data. Hence, it may limit the potential of improving the performance in terms of segmentation accuracy. Moreover, training a GAN that produce

high sample variability requires a large dataset also with a high variability, and, in most of the cases, with corresponding annotations. Considering the data scarcity challenge in the cancer imaging domain, this can be difficult to achieve.

- In some cases, using GANs could be excessive, considering the difficulties related to convergence of competing generator and discriminator parts of the GAN architectures. For example, the recently proposed SynthSeg model (Billot et al., 2020) is based on Gaussian Mixture Models to generate images and train a contrast agnostic segmentation model. Such approaches can be considered as an alternative to avoid common pitfalls of the GAN training process (e.g., mode collapse). However, this approach needs to be further investigated for cancer imaging tasks where the heterogeneity of tumours is challenging.
- A great potential for using synthetic cancer images is to generate common shareable datasets as benchmarks for automated segmentation methods (Bi et al., 2019). Although this benchmark dataset needs its own validation, it can be beneficial in testing the limits of automated methods with systematically controlled test cases. Such benchmark datasets can be generated by controlling the shape, location, size, intensities of tumours, and can simulating diverse images of different domains that reflect the distributions from real institutions. To avoid learning patterns that are only available in synthetic datasets (e.g., checkerboard artifacts), it is a prospect to investigate further metrics that measure the distance of such synthetic datasets to real-world datasets and the generalisation and extrapolation capabilities of models trained on synthetic benchmarks to real-world data.

4.4. Detection and diagnosis challenges

4.4.1. Common issues in diagnosing malignancies

Clinicians' high diagnostic error rates. Studies of radiological error report high ranges of diagnostic error rates (e.g., discordant interpretations in 31%–37% in Oncologic CT, 13% major discrepancies in Neuro CT and MRI) (Brady, 2017). After (McCreddie and Oliver, 2009) critically reviewed the radiology cases of the last 30 months in their clinical centre, they found that from 256 detected errors (62% CT, 12% Ultrasound, 11% MRI, 9% Radiography, 5% Fluoroscopy) in 222 patients, 225 errors (88%) were attributable to poor image interpretation (14 false positive, 155 false negative, 56 misclassifications). A recent literature review on diagnostic errors by Newman-Toker et al. (2021) estimated a false negative rate,²⁰ of 22.5% for lung cancer, 8.9% for breast cancer, 9.6% for colorectal cancer, 2.4% for prostate and 13.6% for melanoma. These findings exemplify the uncomfortably high diagnostic and image interpretation error rates that persist in the field of radiology despite decades of interventions and research (Itri et al., 2018).

The challenge of reducing clinicians' high workload. In some settings, radiologists must interpret one CT or MRI image every 3–4 s in an average 8-h workday (McDonald et al., 2015). Automated CADE and CADx systems can provide a more balanced quality-focused workload for radiologists, where radiologists focus on scrutinising the automated detected lesions (false positive reduction) and areas/patches with high predictive uncertainty (false negative reduction). A benefit of CADE/CADx deep learning models are their real-time inference and strong pattern recognition capabilities that are not readily susceptible to cognitive bias (discussed in 4.3.1), environmental factors (Itri et al., 2018), or inter-observer variability (discussed in 4.3.1).

²⁰ In Newman-Toker et al. (2021) the false negative rates includes both missed (patient encounters at which the diagnosis might have been made but was not) and delayed diagnosis (diagnostic delay relative to urgency of illness detection).

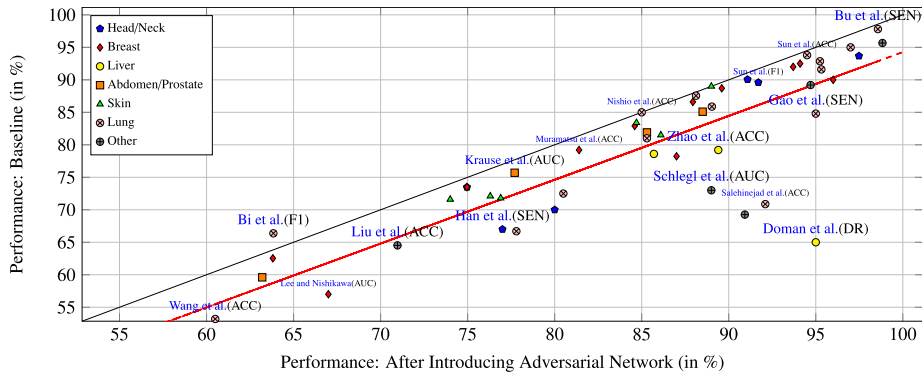


Fig. 12. Scatter plot illustrating the performance improvement attributable to adversarial networks for the surveyed disease diagnosis and detection publications. The shown performance is based on the respective publication's metric reported in Table 5 and include f1-score (F1), sensitivity (SEN), accuracy (ACC), area under the receiver operating characteristic curve (AUC), and detection rate (DR). Each publication is represented by a marker with a colour and shape encoding depicting the publication's anatomical category. The black identity line indicates no change between baseline and adversarial network intervention, while dots below this line represent an improvement. The red regression line depicts the trend of improvement across publications. The author names of a few publications have been manually selected for highlighting based on the distance to the trend line.

Detection model performance on critical edge cases. Challenging cancer imaging problems are the high intra- and inter-tumour heterogeneity (Bi et al., 2019), the detection of small lesions and metastasis across the body (e.g., lymph node involvement and distant metastasis Hosny et al., 2018) and the accurate distinction between malignant and benign tumours (e.g., for detected lung nodules that seem similar on CT scans Hosny et al., 2018). Methods are needed to extend on and further increase the current performance of deep learning detection models (Bi et al., 2019).

4.4.2. GAN applications for cancer detection and diagnosis

As we detail in the following, the capability of adversarial learning to improve malignancy detection has been demonstrated for multiple tumour types and imaging modalities. To this end, Table 5 summarises the collection of recent publications that utilise GANs and adversarial training for cancer detection, classification, and diagnosis.

Figs. 12 and 13 visualise the publications' performance metric values reported in Table 5. Fig. 12 provides visual estimate of the effectiveness of GANs and adversarial training in increasing downstream task performance. A performance increase of approximately 5 percentage points can be observed by comparing the figure's black identity line with the red trend line over publications. We note that no visual pattern seems to be observable that indicates a difference in performance gain between anatomical categories. Across all publications, the performance gains seem not to be a function of the strength of the baseline, as they remain approximately constant with increasing baseline performance. This is indicated by the minimal change of distance between the black identity line and the red trend line throughout the graph. Fig. 13 shows the GAN-induced variation in performance for the years 2017 to 2021 with multiple adversarial models achieving a performance increase of over 10% and most models over 3% on their respective diagnostic downstream task. As emphasised in Section 4.3.2, conclusion drawn from Figs. 12 and 13 have to take publication bias into account. Further, benchmarking and comparison of individual publications based on the presented data in these figures is not part of their intended use due to the differences in baselines, modalities, organs, train and test datasets, and publication objectives.

Adversarial anomaly and outlier detection examples. Schlegl et al. (2017) captured imaging markers relevant for disease prediction using a deep convolutional GAN named AnoGAN. AnoGAN learnt a manifold of normal anatomical variability, accompanying a novel anomaly scoring scheme based on the mapping from image space to a latent space. While Schlegl et al. validated their model on retina optical coherence

tomography images, their unsupervised anomaly detection approach is applicable to other domains including cancer detection, as indicated in Fig. 1(l). Chen et al. (2018b) used a Variational Autoencoder GAN for unsupervised outlier detection using T1 and T2 weighted brain MRI images. The scans from healthy subjects were used to train the auto-encoder model to learn the distribution of healthy images and detect pathological images as outliers. Creswell et al. (2018) proposed a semi-supervised Denoising Adversarial Autoencoder (ssDAAE) to learn a representation based on unlabelled skin lesion images. The semi-supervised part of their CNN-based architecture corresponds to malignancy classification of labelled skin lesions based on the encoded representations of the pretrained DAAE. As the amount of labelled data is smaller than the unlabelled data, the labelled data is used to fine-tune classifier and encoder. In ssDAAE, not only the adversarial autoencoder's chosen prior distribution (Makhzani et al., 2015), but also the class label distribution is discriminated by a discriminator, the latter distinguishing between predicted continuous labels and real binary (malignant/benign) labels. Kuang et al. (2020) applied unsupervised learning to distinguish between benign and malignant lung nodules. In their multi-discriminator GAN (MDGAN) various discriminators scrutinise the realism of generated lung nodule images. After GAN pretraining, an encoder is added in front of the generator to the end-to-end architecture to learn the feature distribution of benign pulmonary nodule images and to map these features into latent space. The benign and malignant lung nodules were scored similarly as in the f-AnoGAN framework (Schlegl et al., 2019), computing and combining an image reconstruction loss and a feature matching loss, the latter comparing the discriminators' feature representations between real and encoded-generated images from intermediate discriminator layers. As exemplified in Fig. 9(g), the model yielded high anomaly scores on malignant images and low anomaly scores on benign images despite limited dataset size. Benson and Beets-Tan (2020) used GANs trained from multi-modal MRI images as a 3-channel input (T1-T2 weighted, FLAIR, ADC MRI) in brain anomaly detection. The training of the generative network was performed using only healthy images together with pseudo-random irregular masks. Despite the training dataset consisting of only 20 subjects, the resulting model increased the anomaly detection rate.

Synthetic detection model training data. Among the GAN publications trying to improve classification and detection performance, data augmentation is the most recurrent approach to balance, vary, and increase the detection model's training set size, as suggested in Fig. 1(k).

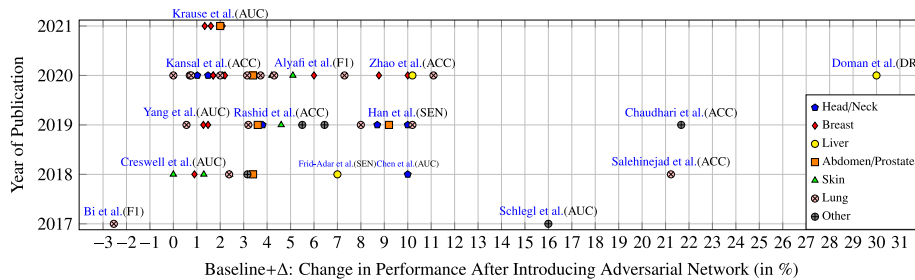


Fig. 13. Scatter plot displaying year of publication and change in performance between baseline and adversarial network method (in %). As in Table 5, the underlying performance metrics include f1-score (F1), sensitivity (SEN), accuracy (ACC), area under the receiver operating characteristic curve (AUC), and detection rate (DR). Each publication is represented by a marker with a colour and shape encoding depicting the publication's anatomical category. Author names have been manually selected at random for highlighting.

For instance in breast imaging, Wu et al. (2018a) trained a class-conditional GAN to perform contextual in-filling to synthesise lesions in healthy scanned mammograms. Guan and Loew (2019) trained a GAN on the same dataset (Heath et al., 2001) to generate synthetic patches with benign and malignant tumours. The synthetic generated patches had clear artifacts and did not match the original dataset distribution. Jendele et al. (2019) used a CycleGAN (Zhu et al., 2017) and both film scanned and digital mammograms to improve binary (malignant/benign) lesion detection using data augmentation. Detecting mammographically-occult breast cancers is another challenging topic addressed by GANs. For instance, Lee and Nishikawa (2020) exploit asymmetries between mammograms of the left and right breasts as signals for finding mammography-occult cancer. They trained a image-conditioned GAN (pix2pix) to generate a healthy synthetic mammogram image of the contralateral breast (e.g., left breast) given the corresponding single-sided mammogram (e.g., right breast) as input. The authors showed that there is a higher similarity (MSE, 2D-correlation) between simulated-real (SR) mammogram pairs than real-real (RR) mammogram pairs in the presence of mammography-occult cancer. Consequently, distinguishing between healthy and mammography-occult mammograms, their classifier yielded a higher performance when trained with both RR and SR similarity as input (AUC = 0.67) than when trained only with RR pair similarity as input (AUC = 0.57). 3-dimensional conditional image synthesis with GANs has been shown, for instance, by Han et al. (2019a), who proposed a 3D Multi-Conditional GAN (3DMCGAN) to generate realistic and diverse nodules placed naturally on lung CT images to boost sensitivity in 3D object detection. Bu et al. (2020) built a 3D image-conditioned GAN based on pix2pix, where the input is a 3D volume of interest (VOI) that is cropped from a lung CT scan and contains a missing region in its centre. Both generator and discriminator contain squeeze-and-excitation (Hu et al., 2018a) residual (He et al., 2016) neural network (SE-ResNet) modules to improve the quality of the synthesised lung nodules. Another example based on lung CT images is the method by Nishio et al. (2020), where the proposed GAN model used masked 3D CT images and nodule size information to generate images.

As to multi-modal training data synthesis, Van Tulder and de Bruijne (2015) replaced missing sequences of a multi-sequence MRI with synthetic data. The authors illustrated that if the synthetic data generation model is more flexible than the classification model, the synthetic data can provide features that the classifier has not extracted from the original data, which can improve the performance. During colonoscopy, depth maps can enable navigation alongside aiding detection and size measurements of polyps. For this reason, Rau et al. (2019) demonstrated the synthesis of depth maps using a image-conditioned GAN (pix2pix) with monocular endoscopic images as input, reporting promising results on synthetic, phantom and real datasets. In breast cancer detection, Muramatsu et al. (2020) translated lesions from lung

CT to breast MMG using cycleGAN yielding a performance improvement in breast mass classification when training a classifier with the domain-translated generated samples.

4.4.3. Future prospects for cancer detection and diagnosis

Granular class distinctions for synthetic tumour images. Further research opportunity exists in exploring a more fine-grained classification of tumours that characterises different subtypes and disease grades instead of binary malignant–benign classification. Being able to robustly distinguish between different disease subtypes with similar imaging phenotypes (e.g., glioblastoma versus primary central nervous system lymphoma Kang et al., 2018) addresses the challenge of reducing diagnostic ambiguity (Bi et al., 2019). GANs can be explored to augment training data with samples of specific tumour subtypes to improve the distinction capabilities of disease detection models. This can be achieved by training a detection model on training data generated by various GANs, where each GAN is trained on a different tumour subtype distribution. Another option we estimate worth exploring is to use the tumour subtype or the disease grade (e.g., the Gleason Score for prostate cancer Hu et al., 2018b) as a conditional input into the GAN to generate additional labelled synthetic training data.

Cancer image interpretation and risk estimation. Besides the detection of prospectively cancerous characteristics in medical scans, ensuring a high accuracy in the subsequent interpretation of these findings are a further challenge in cancer imaging. Improving the interpretation accuracy can reduce the number of unnecessary biopsies and harmful treatments (e.g., mastectomy, radiation therapy, chemotherapy) of indolent tumours (Bi et al., 2019). For instance, the rate of overdiagnosis of non-clinically significant prostate cancer ranges widely between 1.7% up to a noteworthy 67% (Loeb et al., 2014). To address this, detection models can be extended to provide risk and clinical significance estimations. For example, given both an input image, and an array of risk factors (e.g., BRCA1/BRCA2 status for breast cancer Li et al., 2017, comorbidity risks), a deep learning model can weight and evaluate a patient's risk based on learned associations between risk factors and input image features. The GAN framework is an example of this, where clinical, non-clinical and imaging data can be combined, either as conditional input for image generation or as prediction targets. For instance, given an input image, an AC-GAN (Odena et al., 2017; Kapil et al., 2018) can classify the risk as continuous label (see Fig. 9(e)) or, alternatively, a discriminator can be used to assess whether a risk estimate provided by a generator is realistic. Also, a generator can learn a function for transforming and normalising an input image given one or several conditional input target risk factors or tumour characteristics (e.g., a specific mutation status, a present comorbidity, etc.) to generate labelled synthetic training data.

4.5. Treatment and monitoring challenges

After a tumour is detected and properly described, new challenges arise related to planning and execution of medical intervention. In this section we examine these challenges, in particular: tumour profiling and prognosis; challenges related to choice, response and discovery of treatments; as well as further disease monitoring. Table 6 provides an overview of the cancer imaging GANs that are applied to treatment and monitoring challenges, which are discussed in the following.

4.5.1. Disease prognosis and tumour profiling

Challenges for disease prognosis. An accurate prognosis is crucial to plan suitable treatments for cancer patients. However, in specific cases, it could be more beneficial to actively monitor the tumours instead of treating them Bi et al. (2019). Challenges in cancer prognosis include the differentiation between long-term and short term survivors (Bi et al., 2019), patient risk estimation considering the complex intra-tumour heterogeneity of the tumour microenvironment (TME) (Nearchou et al., 2021), or the estimation of the probability of disease stages and tumour growth patterns, which can strongly affect outcome probabilities (Bi et al., 2019). In this sense, GANs (Li et al., 2021b; Kim et al., 2018b) and AI models in general (Cuocolo et al., 2020; Dimitriou et al., 2018) have shown potential in prognosis and survival prediction for oncology patients.

GAN disease prognosis examples. Li et al. (2021b) (in Table 2) show that their GAN-based CT normalisation framework for overcoming the domain shift between images from different centres significantly improves accuracy of classification between short-term and long-term survivors. Ahmed et al. (2021) trained omicsGAN to translate between microRNA and mRNA expression data pairs, but could be readily enhanced to also translate between cancer imaging features and genetic information. The authors evaluate omicsGAN on breast and ovarian cancer datasets and report improved prediction signals for synthetic data tested via cancer outcome classification. Another non-imaging approach is provided by Kim et al. (2018b), who apply a GAN for patient cancer prognosis prediction based on identification of prognostic biomarker genes. They train their GAN on reconstructed human biology pathways data, which allows for highlighting genes relevant to cancer development, resulting in improvement of the prognosis prediction accuracy. In regard to these works on non-imaging approaches, we promote future extensions combining prognostic biomarker genes and -omics data with the phenotypic information present in cancer images into multi-modal prognosis models.

GAN tumour profiling examples. Related to Fig. 1(l), Vu et al. (2020a) propose that image-conditioned GANs (pix2pix) can learn latent characteristics of tissues of tumours that correlate with specific tumour grade. The authors show that when inferring their proposed BenignGAN on malignant tumour tissue images after training it exclusively on benign ones, it generates less realistic results. This allows for quantitative measurement of the differences between the original and the generated image, whereby these differences can be interpreted as tumour grade.

Kapil et al. (2018) explore AC-GAN (Odena et al., 2017) on digital pathology imagery for semi-supervised quantification of the Non-Small-Cell-Lung-Cancer biomarker *programmed death ligand 1* (PD-L1). Their class-conditional generator receives a one-hot encoded PD-L1 label as input to generate a respective biopsy tissue image, while their discriminator receives the image and predicts both PD-L1 label and whether the image is fake or real. The AC-GAN method compares favourably to other supervised and non-generative semi-supervised approaches, and also systematically yields high agreement with visual²¹ tumour proportional scoring (TPS).

²¹ A visual estimation of pathologists of the tumour cell percentage showing PD-L1 staining.

As for the analysis of the TME, Quiros et al. (2019) propose PathologyGAN, which they train on breast and colorectal cancer tissue imagery. This allows for learning the most important tissue phenotype descriptions, and provides a continuous latent representation space, enabling quantification and profiling of differences and similarities between different tumours' tissues. Quiros et al. (2019) show that lesions encoded in an GAN's latent space enable using vector distance measures to find similar lesions that are close in the latent space within large patient cohorts. We highlight the research potential in lesion latent space representations to assess inter-tumour heterogeneity. Also, the treatment strategies and successes of patients with a similar lesion can inform the decision-making process of selecting treatments for a lesion at hand, as denoted by Fig. 1(m).

Outlook on genotypic tumour profiling with phenotypic data. A further challenge is that targeted oncological therapies require genomic and immunological tumour profiling (Cuocolo et al., 2020) and effective linking of tumour genotype and phenotype. Biopsies only allow to analyse the biopsied portion of the tumour's genotype, while also increasing patient risk due to the possibility of dislodging and seeding of neoplastic altered cells (Shyamala et al., 2014; Parmar et al., 2015). Therefore, a trade-off²² exists between minimising the number of biopsies and maximising the biopsy-based information about a tumour's genotype. These reasons and the fact that current methods are invasive, expensive, and time-consuming (Cuocolo et al., 2020) make genotypic tumour profiling an important issue to be addressed by AI cancer imaging methods. In particular adversarial deep learning models are promising to generate the non-biopsied portion of a tumour's genotype after being trained on paired genotype and radiology imaging data.²³ We recommend future studies to explore this line of research, which is regarded as a key challenge for AI in cancer imaging (Bi et al., 2019; Parmar et al., 2015).

4.5.2. Treatment planning and response prediction

Challenges for cancer treatment predictions. A considerable number of malignancies and tumour stages have various possible treatment options and almost no head-to-head evidence to compare them to. Due to that, oncologists need to subjectively select an approved therapy based on their individual experience and exposure (Troyanskaya et al., 2020).

Furthermore, despite existing treatment response assessment frameworks in oncology, inter- and intra-observer variability regarding choice and measurement of target lesions exists among oncologists and radiologists (Levy and Rubin, 2008). To achieve consistency and accuracy in standardised treatment response reporting frameworks (Levy and Rubin, 2008), AI and GAN methods can identify quantitative biomarkers²⁴ from medical images in a reproducible manner useful for risk and treatment response predicts (Hosny et al., 2018).

Apart from the treatment response assessment, treatment response prediction is also challenging, particularly for cancer treatments such as immunotherapy (Bi et al., 2019). In cancer immunogenomics, for instance, unsolved challenges comprise the integration of multi-modal data (e.g., radiomic and genomic biomarkers Bi et al., 2019), immunogenicity prediction for neoantigens, and the longitudinal non-invasive monitoring of the therapy response (Troyanskaya et al., 2020). In regard to the sustainability of a therapy, the inter- and intra-tumour heterogeneity (e.g., in size, shape, morphology, kinetics, texture, etiology) and potential sub-clone treatment survival complicates individual treatment prediction, selection, and response interpretation (Bi et al., 2019).

²² Due to this and due to the high intra-tumour heterogeneity, available biopsy data likely only describes a subset of tumour's clonal cell population.

²³ Imaging data on which the entire lesion is visible to allow learning correlations between phenotypic tumour manifestations and genotype signatures.

²⁴ For example, characteristics and density variations of the parenchyma patterns on breast images (Bi et al., 2019).

Table 6

Overview of adversarially-trained models applied to **treatment and monitoring** challenges. Publications are clustered by section and ordered by year in ascending order.

Publication	Method	Dataset	Modality	Task	Highlights
Disease prognosis					
Kim et al. (2018b)	GAN-based	TCGA (Tomczak et al., 2015), Reactome (Croft et al., 2014; Fabregat et al., 2017)	[non-imaging] multi-omics cancer data	Data synthesis	Biomarker gene identification for pancreas, breast, kidney, brain, and stomach cancer with GANs and PageRank.
Ahmed et al. (2021)	omicsGAN	TCGA (Network et al., 2011; Ciriello et al., 2015)	[non-imaging] ovarian/breast gene expression	Paired translation	microRNA to mRNA translation and vice versa. Synthetic data improves cancer outcome classification.
Tumour profiling					
Kapil et al. (2018)	AC-GAN	Private	Lung histopathology	Classification	AC-GAN CLF of PD-L1 levels for lung tumour tissue images obtained via needle biopsies.
Quiros et al. (2019) PathologyGAN	PathologyGAN	VGH/NKI (Beck et al., 2011), NCT (Kather et al., 2018)	Breast/colorectal histopathology	Representation learning	Learning tissue phenotype descriptions & tumour representations. Combines BigGAN (Brock et al., 2018), StyleGAN (Karras et al., 2019) & RAD (Jolicœur-Martineau, 2018)
Vu et al. (2020a)	BenignGAN	Private	Colorectal histopathology	Paired translation	Edge map-to-image. As trained on only benign, malignant images quantifiable via lower realism.
Treatment response prediction					
Kadurin et al. (2017a) and Kadurin et al. (2017b)	AAE-based druGAN	Pubchem BioAssay (Wang et al., 2014)	[non-imaging] Molecular fingerprint data	Representation learning	AAE for anti-cancer agent drug discovery. AAE input/output: molecular fingerprints & log concentration vectors.
Goldsborough et al. (2017)	CytoGAN	BBBC021 (Ljosa et al., 2012b)	Cytopathology	Representation learning	Grouping cells with similar treatment response via cell image representations. Based on DCGAN, LSGAN, WGAN.
Yoon et al. (2018)	GANITE	USA 89-91 Twins (Almond et al., 2004)	[non-imaging] individualised treatment effects	Multi-class-conditional synthesis	cGANs for individual treatment effect prediction, including unseen counterfactual outcomes and confidence intervals.
Ge et al. (2020)	MGANITE	AML clinical trial (Kornblau et al., 2009)	[non-imaging] individualised treatment effects	Multi-class-conditional synthesis	GANITE extension introducing dosage quantification and continuous and categorical treatment effect estimation.
Bica et al. (2020)	SCIGAN	CGA (Weinstein et al., 2013), MIMIC III (Johnson et al., 2016)	[non-imaging] individualised treatment effects	Multi-class-conditional synthesis	GANITE extension introducing continuous interventions and theoretical explanation for GAN counterfactuals.
Radiation dose planning					
Mahmood et al. (2018)	pix2pix-based	Private	Oropharyngeal CT	Paired translation	Translating CT to 3D dose distributions without requiring hand-crafted features.
Maspero et al. (2018)	pix2pix	Private	Prostate/rectal/cervical CT/MRI	Paired translation	MR-to-CT translation for MR-based radiation dose planning without CT acquisition.
Murakami et al. (2020)	pix2pix	Private	Prostate CT	Paired translation	CT-to-radiation dose distribution image translation without time-consuming contour/organs at risk (OARs) data.
Peng et al. (2020)	pix2pix, CycleGAN	Private	Nasopharyngeal CT/MRI	Unpaired/Paired translation	Comparison of pix2pix & CycleGAN-based generation of CT from MR for radiation dose planning.
Kearney et al. (2020a)	DoseGAN	Private	Prostate CT/PTV/OARs	Paired translation	Synthesis of volumetric dosimetry from CT+PTV+OARs even in the presence of diverse patient anatomy.
Disease tracking & monitoring					
Kim et al. (2019b)	CycleGAN	Private	Liver MRI/CT/dose	Unpaired translation	Pre-treatment MR+CT+dose translation to post-treatment MRI → predicting hepatocellular carcinoma progression.
Elazab et al. (2020)	GP-GAN	BRATS 2014 (Menze et al., 2014)	Cranial MRI	Paired translation	3D U-Net G generating progression image from longitudinal MRI to predict glioma growth between time-step.
Li et al. (2020a)	DC-AL GAN, DCGAN	Private	Cranial MRI	Noise-to-image synthesis	CLF uses D representations to distinguish pseudo- and true glioblastoma progression.

GAN treatment effect estimation examples. In line with Fig. 1(n), Yoon et al. (2018) propose the conditional GAN framework ‘GANITE’, where individual treatment effect prediction allows for accounting for unseen, counterfactual outcomes of treatment. GANITE consists of two GANs: first, a counterfactual GAN is trained on feature and treatment vectors along with the factual outcome data. Then, the trained generator’s output is used for creating a dataset, on which the other GAN, called ITE (Individual Treatment Response) GAN, is being trained. GANITE provides confidence intervals along with the prediction, while being readily scalable for any number of treatments. However, it does not

allow for taking time, dosage or other treatment parameters into account. MGANITE, proposed by Ge et al. (2020), extends GANITE by introducing dosage quantification, and thus enables continuous and categorical treatment effect estimations. SCIGAN (Bica et al., 2020) also extends upon GANITE and predicts outcomes of continuous rather than one-time interventions and the authors further provide theoretical justification for GANs’ success in learning counterfactual outcomes. As to the problem of individual treatment response prediction, we suggest that quantitative comparisons of GAN-generated expected post-treatment images with real post-treatment images can yield interesting insight for tumour interpretation. We encourage future work to explore

generating such post-treatment tumour images given a treatment parameter and a pre-treatment tumour image as conditional inputs. With varying treatment parameters as input, it is to be investigated whether GANs can inform treatment selection by simulating various treatment scenarios prior to treatment allocation or whether GANs can help to understand and evaluate treatment effects by generating counterfactual outcome images after treatment application.

Goldsborough et al. (2017) present an approach called CytoGAN, where they synthesise fluorescence microscopy cell images using DC-GAN, LSGAN, or WGAN. The discriminator's latent representations learnt during synthesis enable grouping encoded cell images together that have similar cellular reactions to treatment by chemicals of known classes (morphological profiling).²⁵ Even though the authors reported that CytoGAN obtained inferior result²⁶ compared to classical, widely applied methods such as CellProfiler (Singh et al., 2014), using GANs to group tumour cells representations to inform chemical cancer treatment allocation decisions is an interesting approach in the realm of treatment selection, development (Kadurin et al., 2017a,b) and response prediction.

GAN radiation dose planning examples. As radiation therapy planning is labour-intensive and time-consuming, researchers have been spurred to pursue automated planning processes (Sharpe et al., 2014). As outlined in the following and suggested by Fig. 1(o), the challenge of automated radiation therapy planning can be approached using GANs.

By framing radiation dose planning as an image colourisation problem, Mahmood et al. (2018) introduced an end-to-end GAN-based solution, which predicts 3D radiation dose distributions from CT without the requirement of hand-crafted features. They trained their model on Oropharyngeal cancer data along with three traditional ML models and a standard CNN as baselines. The authors trained a pix2pix (Isola et al., 2017) GAN on 2D CT imagery, and then fed the generated dose distributions to an inverse optimisation (IO) model (Babier et al., 2018), in order to generate optimised plans. Their evaluation showed that their GAN plans outperformed the baseline methods in all clinical metrics.

Kazemifar et al. (2020) (in Table 2) proposed a cGAN with U-Net generator for paired MRI to CT translation. Using conventional dose calculation algorithms, the authors compared the dose computed for real CT and generated CT, where the latter showed high dosimetric accuracy. The study, hence, demonstrates the feasibility of synthetic CT for intensity-modulated proton therapy planning for brain tumour cases, where only MRI scans are available.

Maspero et al. (2018) proposed a GAN-assisted approach to quicken the process of MR-based radiation dose planning, by using a pix2pix for generating synthetic CTs (sCTs) required for this task. They show that a conditional GAN trained on prostate cancer patient data can successfully generate sCTs of the entire pelvis.

A similar task has also been addressed by Peng et al. (2020). Their work compares two GAN approaches: one is based on pix2pix and the other on a CycleGAN (Zhu et al., 2017). The main difference between these two approaches was that pix2pix was trained using registered MR-CT pairs of images, whereas CycleGAN was trained on unregistered pairs. Ultimately, the authors report pix2pix to achieve results (i.e. mean absolute error) superior to CycleGAN, and highlight difficulties in generating high-density bony tissues using CycleGAN.

The recently introduced attention-aware DoseGAN (Kearney et al., 2020a) overcomes the challenges of volumetric dose prediction in the presence of diverse patient anatomy. As illustrated in Fig. 14, DoseGAN is based on a variation of the pix2pix architecture with a 3D encoder-decoder generator (L1 loss) and a patch-based patch-GAN discriminator (adversarial loss). The generator was trained on concatenated CT, planning target volume (PTV) and organs at risk

(OARs) data of prostate cancer patients, and the discriminator's objective was to distinguish the real dose volumes from the generated ones. Both qualitatively and quantitatively, DoseGAN was able to synthesise more realistic volumetric doses compared to current alternative state-of-the-art methods.

Murakami et al. (2020) published another GAN-based fully automated approach to dose distribution of Intensity-Modulated Radiation Therapy (IMRT) for prostate cancer. The novelty of their solution is that it does not require the tumour contour information, which is time-consuming to create, to successfully predict the dose based on the given CT dataset. Their approach consists of two pix2pix-based architectures, one trained on paired CT and radiation dose distribution images, and the other trained on paired structure images and radiation dose distribution images. From the generated radiation dose distribution images the dosimetric parameters for the PTV and OARs are computed. The generated dosimetric parameters differed on average only between 1%–3% with respect to the original ground truth dosimetric parameters.

Koike et al. (2020) proposed a CycleGAN for dose estimation for head and neck CT images with metal artifact removal in CT-to-CT image translation as described in Table 2. Providing consistent dose calculation against metal artifacts for head and neck IMRT, their approach achieves dose calculation performance similar to commercial metal artifact removal methods.

4.5.3. Disease tracking and monitoring

Challenges in tracking and modelling tumour progression. Tumour progression is challenging to model (Huang et al., 2020) and commonly requires rich, multi-modal longitudinal data sets. As cancerous cells acquire growth advantages through genetic mutation in a process arguably analogous to Darwinian evolution (Hanahan and Weinberg, 2000), it is difficult to predict which of the many sub-clones in the TME will outgrow the other clones. A tumour lesion is, hence, constantly evolving in phenotype and genotype (Bi et al., 2019) and might acquire dangerous further mutations over time, anytime. The TME's respective impact is exemplified by the stage II colorectal cancer outcome classification performance gain in Dimitriou et al. (2018), which is likely attributable to the high prognostic value of the TME information in their training data.

In addition, concurrent conditions and alterations in the organ system surrounding a tumour, but also in distant organs may not only remain undetected, but could also influence patient health and progression (Bi et al., 2019). GANs can generate hypothetical comorbidity data²⁷ to aid awareness, testing, finding, and analysis of complex disease and comorbidity patterns. A further difficulty for tumour progression modelling is the a priori unknown effect of treatment. Treatment effects may even remain partly unknown after treatment for example in the case of radiation therapy²⁸ (Verma et al., 2013) or after surgery²⁹ (Bi et al., 2019).

GAN tumour progression modelling examples. Relating to Fig. 1(p), GANs can not only diversify the training data, but can also be applied to simulate and explore disease progression scenarios (Elazab et al., 2020). For instance, Elazab et al. (2020) propose GP-GAN, which uses stacked 3D conditional GANs for growth prediction of glioma based on longitudinal MR images. The generator is based on the U-Net architecture (Ronneberger et al., 2015) and the segmented feature maps

²⁷ For example from EHR (Hwang et al., 2017; Dashtban and Li, 2020), imaging data, or a combination thereof.

²⁸ Radiation therapy can result in destruction of the normal tissue (e.g., radionecrosis) surrounding the tumour. Such heterogeneous normal tissue can become difficult to characterise and distinguish from the cancerous tissue (Verma et al., 2013).

²⁹ It is challenging to quantify the volume of remaining tumour residuals after surgical removal (Bi et al., 2019).

²⁵ CytoGAN uses an approach comparable to the one shown in Fig. 9(g).

²⁶ i.e. mechanism-of-action classification accuracy.

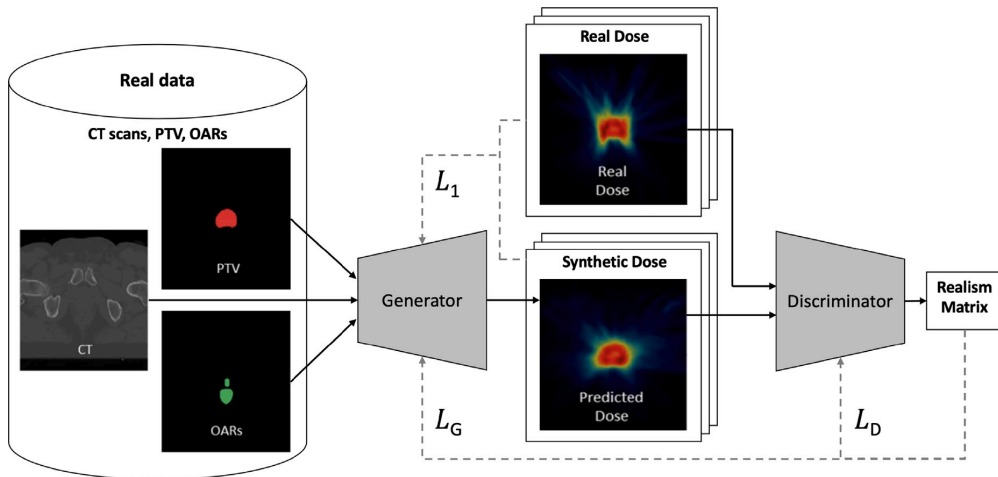


Fig. 14. GAN architecture of DoseGAN adapted from Kearney et al. (2020a) and based on pix2pix (Isola et al., 2017). Given concatenated CT scans, planning target volume (PTV) and organs at risk (OARs), the generator of DoseGAN addresses the challenge of volumetric dose prediction for prostate cancer patients.

are used in the training process. Kim et al. (2019b) trained a CycleGAN on concatenated pre-treatment MR, CT and dose images (i.e. resulting in one 3-channel image) of patients with hepatocellular carcinoma to generate follow-up enhanced MR images. This enables tumour image progression prediction after radiation treatment, whereby CycleGAN outperformed a vanilla GAN baseline.

The deep convolutional (DC) (Radford et al., 2015) - AlexNet (AL) (Krizhevsky et al., 2012) GAN (DC-AL GAN) proposed by Li et al. (2020a) is trained on longitudinal diffusion tensor imaging (DTI) data of pseudoprogression (PsP) and true tumour progression (TTP) in glioblastoma multiforme (GBM) patients. Both of these progression types can occur after standard treatment³⁰ and they are often difficult to differentiate due to similarities in shape and intensity. In DC-AL GAN, representations are extracted from various layers of its AlexNet discriminator that is trained on discriminating between real and generated DTI images. These representations are then used to train a support vector machine (SVM) classifier to distinguish between PsP and TTP samples achieving promising performance.

We recommend further studies to extend on these first adversarial learning disease progression modelling approaches. One potential research direction are GANs that simulate environment and tumour dependent progression patterns based on conditional input data such as the tumour's gene expression data (Xu et al., 2020) or the progressed time between original image and generated progression image (e.g., time passed between image acquisitions or since treatment exposure). To this end, unexpected changes of a tumour may be uncovered between time points or deviations from a tumour's biopsy proven genotypic growth expectations.³¹

5. Meta-analysis: Towards a framework for the assessment of trustworthiness and validation

5.1. Trustworthiness of medical image synthesis studies

Section 4 presented an extensive analysis of the challenges, existing publications, and state-of-the-art data synthesis and adversarial

network methods in cancer imaging. While the methodologies, experiments, and results of these studies were elaborated, their validity and trustworthiness was not specifically addressed. The validity and trustworthiness varies between studies and depends on the breadth and depth of the methodological evaluation and the analysis of potential limitations. In the absence of a rigorous evaluation indicating otherwise, the methodology and experimental results of a study cannot be readily assumed to be transferable across domains, settings, tasks, datasets and modalities. Hence, while a study reports promising results for a particular task and seemingly solves the task's underlying (cancer imaging) challenge, modest changes in the dataset, evaluation method or evaluation metrics can lead to different results and conclusions. This points to the need of a principled assessment of trustworthiness and validity of studies in the cancer and medical imaging domains, in particular, the ones contributing and evaluating synthetic data and data generation methodology.

Some frameworks have proposed guidelines and best practices for the development of trustworthy artificial intelligence solutions in medical imaging (Lekadir et al., 2021; Hasani et al., 2022). However, to the best of our knowledge, no framework has been proposed for trustworthiness assessment of studies focused on medical image synthesis solutions. Building upon the FUTURE-AI consensus guidelines (Lekadir et al., 2021) and the lesson's learned from the extensive analysis of the 164 publications presented in Section 4, we propose the Synthesis Study Trustworthiness Test (*SynTRUST*) as a principled framework to evaluate medical image synthesis studies.

5.2. Proposing the *SynTRUST* framework

The Synthesis Study Trustworthiness Test (*SynTRUST*) framework consists of a principled set of measures to assess the trustworthiness and validity of studies proposing generative models, synthetic data, or adversarial training methods in medical and cancer imaging. It is based on five core principles, namely,

- (i) *Thoroughness* of experimental design and validation.
- (ii) *Reproducibility* and transparency of results, data, models, and implementation.
- (iii) *Usefulness* and versatility of synthesis method, model, and generated data.
- (iv) *Scalability* and transferability of the methodology and the results across clinical domains.

³⁰ Pseudoprogression occurs in 20%–30% of GBM patients (Li et al., 2020a).

³¹ For example, by comparing the original patient image after progression with the GAN-generated predicted image (or its latent representation) after progression for time spans of interest.

Table 7
Illustration of the *SynTRUST* Framework for evaluation of trustworthiness and validation vigour of studies that propose generative models, synthetic data, or adversarial training in medical and cancer imaging. *SynTRUST* is based on the 5 core principles *Thoroughness*, *Reproducibility*, *Usefulness*, *Scalability*, and *Tenability*. For each overarching principle, a set of concrete corresponding measures is defined. Each of the 26 *SynTRUST* measures is associated to an ID for reference and is assigned an importance rating, where 1 stands for *Essential*, 2 for *Desirable*, and 3 for *Recommended*.

Principle	ID	Rating	Definition
Thoroughness: Validity of experiments			
Minimum test set size	Th1	1	Representative test set size should allow confident conclusions (e.g., > 30 cases, > 100 images, ≥ 20% of training data).
Multi-metric reporting	Th2	1	Multiple standardised metrics (e.g., FID, SSIM, downstream task metric) evaluate the synthesis method (e.g., ≥ 2).
Multiple result validation runs	Th3	2	Mean & variance over multiple runs to be reported for all metrics (e.g., ≥ 3 random seeds, or ≥ 5-fold cross-validation).
Fair baseline comparison	Th4	2	Fairest-possible comparison with closest-possible generative/adversarial/downstream task model baseline.
Principled benchmark definition	Th5	2	Systematic construction of heterogeneous (patients, pathologies, acquisitions) test set(s), without data leaking.
Statistical significance testing	Th6	3	Validation that reported performance variation & improvements are statistically significant.
Ablation study of key components	Th7	3	Testing removal of both downstream & generative model parts for insight on impact.
Effect of varying training set sizes	Th8	3	Testing data scarcity impact by systematically reducing downstream and generative model training data.
Reproducibility: Transparency of study			
Detailed reporting of design decisions	R1	1	Study design & experiments are defined & reported with rationales and attention to detail.
Public availability of dataset	R2	2	At least one reported evaluation dataset is publicly accessible allowing repetition of experiments.
Public availability of software code	R3	2	Source code is shared in publicly accessible repository providing method implementation, ideally with documentation.
Public availability of model weights	R4	3	Sharing of model weights for reusing the trained model for faster and sustainable reproducing of experiments.
Usefulness: Versatility of synthesis method			
Synthesis method usability testing	U1	1	Solid generative/adversarial model usefulness evaluation on at least one community-defined (downstream) clinical task.
Quantitative quality measurement	U2	2	Assessment of synthetic data quality (e.g., via FID) or adversarial loss and their correlation with downstream task metrics.
Qualitative quality measurement	U3	3	Observer study with clinicians to assess synthesis model output on realism, utility, quality, and diversity.
Mode collapse analysis	U4	3	Diversity of generative model modes is analysed (e.g., via visual inspection and t-sne of synthetic & real distributions).
Scalability: Transferability of methodology			
Real-world representing data	S1	1	Evaluation on cases & samples highly representative of medically-relevant real-world clinical data.
Multi-dataset evaluation	S2	2	Evaluation of generative model on multiple datasets/modalities demonstrating scalability, ideally for different organs.
Multi-centre evaluation	S3	3	Evaluation of generative model per centre showing generalisability across centre-specific variations.
Multi-downstream task evaluation	S4	3	Evaluate generative model versatility via test with multiple downstream models & tasks (e.g. segmentation, classification).
Downstream task robustness evaluation	S5	3	Performance variation test for simulated train & test acquisition, manifestation, population, annotation, prevalence shifts.
Tenability: Acceptability of trained model			
Condition adherence testing	Te1	1	Test of preciseness & reliability of presence of (input) conditions in the (synthetic) data (e.g., via classification).
Bias awareness analysis	Te2	2	Discussion & analysis how bias from dataset (e.g., age, gender, ethnicity, in/exclusion criteria) transfers into model.
Model hallucination tendency analysis	Te3	3	Analysis of undesired removal/addition of features such as artifacts or tumours (e.g., via inspection or classification).
Fairness variation testing	Te4	3	Change in fairness is measured for generative/adversarial model intervention (e.g., via equalised odds in downstream task).
Privacy preservation testing	Te5	3	Investigation of patient-identifying feature leakage and training data reconstruction risk given generative model (output).

- (v) *Tenability*, acceptability, and reliability of the properties of the model and respective synthetic data.
- The methodology applied to derive the *SynTRUST* framework is composed of several consecutive steps, outlined as follows.
1. Observation of experimental evaluation methods in the surveyed cancer imaging papers.

2. Questioning to which extent an observed study concludes with a generally-applicable, scientifically-sound finding.

3. Definition of causes as to why the results of the study are limited in general-applicability and trustworthiness.

4. Suggestion of additional validation methods that can increase the study's general-applicability.

5. Grouping and formalisation of suggestions into 26 concrete validation measures.

6. Definition of an overarching principle for each group of measures resulting in the 5 core principles: *Thoroughness*, *Reproducibility*, *Usefulness*, *Scalability*, and *Tenability*.

7. Refinement of the measures to complement with and extend on expert consensus on best practices for the application of artificial intelligence in medical imaging (Lekadir et al., 2021).

8. Importance rating of each measure from 1 to 3 based on their estimated impact on trustworthiness. A rating of 1 indicates *essential* measures with the highest importance, a rating of 2 characterises *desirable* measures, and a rating of 3 depicts measures that are *recommended* additions to a study.

The resulting *SynTRUST* framework is illustrated in Table 7. Table 7 contains the title, the definition, the importance rating, and an ID for reference for each of the 26 measures, grouped by the 5 *SynTRUST* principles.
- 5.3. Analysis of cancer imaging challenges using *SynTRUST*
- 5.3.1. *SynTRUST* study curation
- Towards the objective of evaluating the trustworthiness of cancer imaging solutions, we demonstrate in the following how the *SynTRUST* framework can be used to analyse medical imaging publications. This not only shows the practicability of the *SynTRUST* framework, but also estimates the trustworthiness of current results in the field. The latter allows to corroborate concrete quality-controlled conclusions about the progress and state-of-the-art in adversarial networks in cancer imaging.
- 32

Table 8
Selection of studies that employ data synthesis and adversarial networks methodology curated based on their promising potential towards solving the cancer imaging challenges surveyed in Sections 4.1–4.5. Each of the studies represents one concrete proposed solution to one of the challenges.

Cancer imaging challenge	Proposed solution	Representative study
Imbalanced/biased data (4.1.2)	Adversarially-trained bias-free representations	Li et al. (2021a)
Dataset shifts (4.1.3)	Multi-modal image translation	Yurt et al. (2019)
Uncertain synthetic data usability (4.1.4)	Feature hallucination evaluation metric	Cohen et al. (2018b,a)
Uncurated data (4.1.5)	Generative image correction & denoising model	Armanious et al. (2020)
Privacy risks in data sharing (4.2.1)	Federated (differentially-private) image synthesis	Chang et al. (2020b,a)
Adversarial attacks and defences (4.2.5)	Adversarial example-based augmentation	Liu et al. (2020b)
Costly human annotation (4.3.1)	Uncertainty-aware annotation generation	Hu et al. (2020)
Weak domain generalisation (4.3.2)	Adversarially-trained cross-domain segmentation	Kamnitsas et al. (2017)
Extracted feature variation (4.3.2)	Discriminator learning radiomics correlations	Xiao et al. (2019)
Intra/inter-observer variability (4.3.2)	Observer averaging via mask discriminator	Sarker et al. (2019)
Radiologists' high error rate (4.4.1)	Detection improving synthetic data augmentation	Zhao et al. (2020a)
Intra/inter-tumour heterogeneity (4.4.2)	Adversarially-trained anomaly detection	Kuang et al. (2020)
Uncertain tumour profiles (4.5.1)	Adversarially-trained representation comparison	Quiros et al. (2019)
Unknown treatment response (4.5.2)	Semi-supervised treatment biomarker quantification	Kapil et al. (2018)
Unknown treatment dose (4.5.2)	Synthesis of volumetric dosimetry images	Kearney et al. (2020a)
Uncertain disease progression (4.5.3)	Tumour progression image generation	Elazab et al. (2020)

Table 9
Results of the in-depth analysis of all *essential* and *desirable* measures of the *SynTRUST* framework for studies proposing adversarial network methodology. The analysed studies are selected in Table 8 and represent solutions to key cancer imaging challenges. The *SynTRUST* measures are referenced by ID from Table 7. The blue check mark symbol indicates a positive evaluation, while the red and orange cross symbols respectively indicate a negative evaluation of an essential or desirable measure. The evaluated *essential* measures are *minimum test set size* (Th1), *multi-metric reporting* (Th2), *detailed reporting of design decisions* (R1), *synthesis method usability testing* (U1), *real-world representing data* (S1), *condition adherence testing* (Te1). The evaluated *desirable* measures are *multiple result validation runs* (Th3), *fair baseline comparison* (Th4), *principled benchmark definition* (Th5), *public availability of dataset* (R2), *public availability of software code* (R3), *quantitative quality measurement* (U2), *multi-dataset evaluation* (S2), *bias awareness analysis* (Te2).

Representative study	SynTRUST framework														
	1: Essential measures						2: Desirable measures								
	Th1	Th2	R1	U1	S1	Te1	Th3	Th4	Th5	R2	R3	U2	S2	Te2	
Li et al. (2021a)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	
Yurt et al. (2019)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✗	
Cohen et al. (2018b,a)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	
Armanious et al. (2020)	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	
Chang et al. (2020b,a)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✗	
Liu et al. (2020b)	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	
Hu et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	
Kamnitsas et al. (2017)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗	
Xiao et al. (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	
Sarker et al. (2019)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✗	
Zhao et al. (2020a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	
Kuang et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	
Quiros et al. (2019)	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	
Kapil et al. (2018)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	
Kearney et al. (2020a)	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	✗	
Elazab et al. (2020)	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	

In our analysis we first sample the present-day challenges in cancer imaging that were surveyed in Section 4 and summarised in Fig. 1. Next, we carefully select representative adversarial network publications to represent a particular challenge and its solution. This selection is based on the criteria that the publication (a) proposes a particularly promising solution to its respective challenge, (b) contributes a methodology that is generally-applicable across domains and (c) report promising results. Most of the sampled publications further (d) have shown more impact and were referenced in other relevant studies. The selected studies are displayed in Table 8 together with their representative solution and associated cancer imaging challenge.

5.3.2. *SynTRUST* study assessment

Next, we analyse each of the selected publications independently based on the *SynTRUST* framework. We choose to base our analysis on the most important measures of the *SynTRUST* framework that, as shown in Table 7, have received either a rating of 1 as *essential* or a rating of 2 as *desirable*. For the sake of conciseness, we leave the analysis of less critical measures rated as 3 (*recommended*) to further studies. The results of our analysis of each of the selected publications are summarised in Table 9.

Essential SynTRUST measures. We observe that the analysed studies overall show strong trustworthiness and validity considering the *essential* measures: 11 out of 16 studies fulfil all of the essential criteria, while the remaining 5 studies fulfil all but one essential measures. For 3 out of these 5 studies, the only essential measure that is not fulfilled is Th1 (*minimum test set size*). For instance, studies that pioneer methodologies on promising new clinical applications, such as generative tumour progression modelling (Elazab et al., 2020), it is particularly challenging to encounter datasets suitable for the clinical task at hand. Even though the number of test images exceeds the defined minimum of 100 in these studies, the number of different patients (cases) is lower than 30. 30 was defined as the indicative minimum of cases to allow for conclusions for the larger patient population.³² All 16 studies have a detailed reporting of design decisions (R1), train and test on real-world representing clinical data (S1), and test the conditions of their adversarial network (Te1). Also, 15 out of 16 studies report multiple standardised performance metrics to evaluate the adversarial network (Th2) and demonstrate their method's usefulness on a clinically relevant downstream task (U1). In sum, the result for the essential

³² Based on the central limit theorem, 30 is a popular choice and rule-of-thumb for the minimum sample size of a population.

measures demonstrates that the reported performance and progress of the analysed studies are considerably reliable and trustworthy.

Desirable SynTRUST measures. While the 6 *essential* basic trustworthiness requirements are mostly fulfilled, the result for the 8 *desirable* measures is more varied. This highlights that the studies have a general high level of trustworthiness, but a lower level of trustworthiness for the more specific and nuanced aspects of their reported results and validations. For instance, while 15 out of 16 studies included a comparison with a suitable baseline (Th4), multiple studies did not accomplish a positive evaluation of Th3 (8), Th5 (9), R2 (7), R3 (11), U2 (11), S2 (7), and Te2 (15).

- Regarding Th3, often studies defined a static train and test set without running experiments multiple times. For example, multiple different random seed network weight initialisations or k-fold cross-validation are options to corroborate results by demonstrating stable performance with reported mean and standard deviation across runs/folds.
- Regarding Th5, in general the train-test split ensured no data leaking between training and testing sets, e.g., with images from the same patient not being in both sets. However, the benchmark test sets were often not defined systematically to ensure validating the methods on a varied distribution of, e.g., cases, patients, pathologies, and acquisition parameters.
- For R2, we observe that often the studies' datasets are not public available, which limits the reproducibility of the results. Often, this is due to the collection and usage of private patient data from hospitals. Further limiting factors are the high effort to repeat the study on public datasets or the specificity of the clinical task rendering its evaluation non-viable on the available public datasets.
- Analysing R3 shows that the software implementing the studies' methods and experiments is often not shared publicly in code repositories, which reduces reproducibility and impedes rerunning experiments with exactly the same code base used in the respective study.
- Regarding U2, often the correlation between (a) the downstream tasks and (b) either the synthetic data quality (e.g., in the case of generative models) or the adversarial loss (e.g. in the case of adversarial training) is not analysed. Such an analysis informs on the usefulness of the quality of the respective model and on its contribution to the results on the clinical task.
- As to S2, often the method is validated on, both, (a) a single dataset and (b) a single modality, while a desirable evaluation would use multiple datasets, modalities, ideally further demonstrating the method's transferability across organs, clinical domains and acquisition protocols.
- For Te2, we note the general absence of an analysis of the bias that is transferred from the training dataset into the models. For instance, a model trained on a homogeneous patient population sample, e.g., in terms of gender, sex, ethnicity, geography, likely is biased towards this subset of the overall population and can result in unequal treatment of patients from other subsets. Model biases can be detected by reviewing (a) the dataset statistics, (b) the model performance shifts on carefully subset patient samples, and (c) the exclusion and inclusion criteria applied in the data acquisition and curation processes. This enables to report and potentially mitigate otherwise unknown model biases, which increases the knowledge and reliability of a model's properties.

In concluding our meta-analysis, we highlight the high general level of trustworthiness of the selected adversarial network publications based on our assessment of the *essential SynTRUST* measures. This demonstrates technical maturity of adversarial training and image synthesis methods in cancer imaging. As described in the Sections 4.1–4.5, many approaches towards solving the challenges in cancer imaging

are not yet fully explored. Nonetheless, the solutions that have been pioneered and validated are shown to be relatively trustworthy and solid.

However, our meta-analysis also revealed that specific *desirable* trustworthiness criteria that go beyond basic *essential* validation are often not fulfilled, even by the most promising and in-depth studies in the field. For instance, a wider practice of data and code sharing is desirable. Closing this gap will not only increase reproducibility, but also accelerate adoption of existing methods and further innovation. Apart from that, the validation of biases and fairness criteria in datasets and models is largely overlooked despite its importance to ensure a model's acceptability and trust in the clinical setting.

We motivate further studies to address and build upon the gaps our analysis has revealed regarding the trustworthiness of existing cancer imaging studies. In this regard, we highlight the *SynTRUST* framework not only as a means for study evaluation, but also as a guideline guiding the design of future image synthesis studies.

6. Discussion and future perspectives

6.1. Adversarial methods in cancer imaging over the years

As presented in Fig. 15(c), we have included 164 of the surveyed GAN-based data synthesis and adversarial training publications in the timeframe from 2017 until March 7th 2021. We observe that the numbers of these cancer imaging GAN publications have been increasing from 2017 to 2020 from 10 to 63 with a surprising slight drop between 2018 to 2019 (41 to 38). The final number of respective publications for 2021 is still pending. The trend towards publications that propose GANs and adversarial training to solve cancer imaging challenges demonstrates the considerable research attention that the adversarial learning scheme has been receiving in this field. Following our literature review in Section 4, the need for further research in adversarial networks seems not yet to be met. We were able to highlight various lines of research for GANs and adversarial training in oncology, radiology, and pathology that have received limited research attention or are untapped research potentials. These potentials indicate a continuation of the trend towards more data synthesis and adversarial training applications and standardised integration of GAN-generated synthetic data into medical image analysis pipelines and software solutions.

6.2. Modality biases

In regard to imaging modalities, we analyse in Fig. 15(b) how much research attention each modality has received in terms of the number of corresponding publications. By far, MRI and CT are the most dominant modalities with 61, and 53 publications, respectively, followed by MMG (13), dermoscopy (12) and PET (6). The wide spread between MRI and CT and less investigated domains such as endoscopy (3), ultrasound (3), and digital tomosynthesis (0) is to be critically remarked. Due to variations in the imaging data between these modalities (e.g., spatial resolutions, pixel dimensions, domain shifts), it cannot be readily assumed that a GAN application with desirable results in one modality will produce equally desirable results in another. Due to that and with awareness of the clinical importance of MRI and CT, we suggest a more balanced application of GANs and adversarial training across modalities including experiments on rare modalities to demonstrate the clinical versatility and applicability of GAN-based solutions. Alongside the open-access datasets described by Diaz et al. (2021), we highlight the following additional recent open datasets to facilitate experiments on some of the cancer imaging modalities that we found to be less explored:

- Breast tomosynthesis: BCS-DBT (Buda et al., 2021)
- PET-CT: Lung-PET-CT-Dx (Li et al., 2020b)
- Endoscopy: HyperKvasir (Borgli et al., 2020)
- Dermatology: HAM10000 (Tschandl et al., 2018)
- Cytology: CERVIX93 (Phoulady and Mouton, 2018)
- Thoracic X-ray: Node21 (Soganciglu et al., 2021)

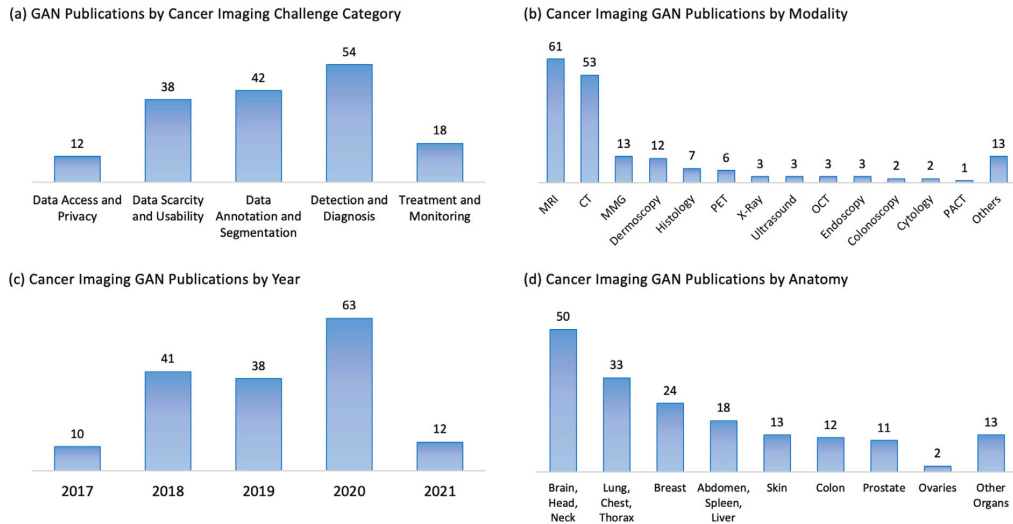


Fig. 15. Histograms showing the distribution of the 164 analysed GAN publications in this paper by (a) cancer imaging challenge category, (b) imaging modality, (c) year of publication, and (d) anatomy/organ. These numbers are retrieved exclusively from the information in Tables 2–6 of the respective Sections 4.1–4.5. Note that (b) and (d) contain more publications in total than (a) and (c), which is caused by GAN publications that evaluate on (and are assigned to) more than one modality (b) and/or anatomy (d) due to multiple experiments or cross-domain translation. In (c), the count for 2021 is not final, as the GAN papers herein analysed have been published on or before 7th March 2021.

6.3. Anatomy biases

In comparison, the GAN-based solutions per anatomy are more evenly spread, but still show a clear trend towards brain, head, neck (50), lung, chest, thorax (33) and breast (24). We suspect these spreads are due to the availability of few well-known widely-used curated benchmark datasets (Menze et al., 2014; Armato III et al., 2011; Heath et al., 2001; Moreira et al., 2012) resulting in underexposure of organs and modalities with less publicly available data resources. Where possible, we recommend evaluating GAN-based data synthesis and adversarial training on a range of different tasks and organs. This can avoid iterating towards non-transferable solutions tuned for specific datasets with limited generalisation capabilities. Said generalisation capabilities are critical for beneficial usage in clinical environments where dynamic data processing requirements and dataset shifts (e.g., multi-vendor, multi-scanner, multi-modal, multi-organ, multi-centre) commonly exist.

6.4. Cancer imaging challenge category biases

Fig. 15(a) displays the distribution of GAN publications across cancer imaging challenge categories that correspond to the subsections of Section 4. While the Sections 4.4 detection and diagnosis (54) and 4.4 data annotation and segmentation (42), and 4.1 data scarcity and usability (38) have received much research attention, Sections 4.5 treatment and monitoring (18) and 4.2 data access and privacy (12) contain substantially less GAN-related publications. This spread can be anticipated considering that classification and segmentation are popular computer vision problems and common objectives in publicly available medical imaging benchmark datasets. Early detected cancerous cells likely have had less time to acquire malignant genetic mutations (Hanahan and Weinberg, 2000, 2011) than their latter detected counterparts, which, by then, might have acquired more treatment-resistant alterations and subclone cell populations. Hence, automated early detection, location and diagnosis can provide high clinical impact via improved cancer treatment prospects, which likely influences the trend towards detection and segmentation-related GAN publications.

6.5. Well-validated adversarial network solutions

Our survey uncovers in Sections 4.1, 4.3, and 4.4 that a vast amount of cancer imaging literature exists around a few common adversarial network solutions.

The most common application of GANs is data augmentation, where synthetic data is added to the training dataset to yield an improved downstream task performance. Such data augmentation can be further used to balance imbalanced datasets, which, for instance, often include much more benign tumour images than malignant ones.

A further well-explored application of GANs is domain adaptation via adversarial training, where a domain-adversarial loss is backpropagated into a downstream task model. Domain mapping is a related application, where images are translated from one domain to another. In general, GANs learn to translate between one source and one target domain. However, promising work has extended this technique to cross-modal synthesis between multiple domains (Yurt et al., 2019; Li et al., 2019a; Zhou et al., 2020), which remains an area with much clinically-relevant research potential. Similarly, GANs for super resolution and data curation including artifact removal and image denoising achieve desirable performance and real-world applicability.

Image-to-image translating GANs can remove or hallucinate features such as tumours (Cohen et al., 2018b,a) into generated images. While this can be a major concern for clinical adoption, it also opens an avenue for future research into automated detection and assessment of removed or hallucinated features and sheds light on the need for additional metrics for GAN condition-adherence and synthetic data evaluation.

Furthermore, we observe that the discriminator and its associated adversarial loss can be flexibly used to classify any type of model output without necessarily following the purpose of data generation. For example, discriminator can predict whether a segmentation mask is real or created by a segmentation model, which enables the model to learn to output more globally coherent segmentation masks.

6.6. New solutions for unexploited areas

Patient privacy. We promote future work on the less researched open challenges in Section 4.2, where we describe the promising research

potential of adversarial networks in patient data privacy and security. We note that secure patient data is required for legal and ethical patient data sharing and usage, which, on the other hand, is required for successful training of state-of-the-art downstream task models. For instance, sharing GANs instead of private patient data can reduce data sharing constraints, while maintaining data utility (Szafranowska et al., 2022). Furthermore, GANs can be trained both in a federated learning setup as well as in a differential-privacy setup. Both of these techniques can be combined to further reduce privacy risks such as the risk of generating synthetic imaging data attributable to a specific patient. Further unexploited research potential lies in adversarial identity obfuscation both on image level, as well as on latent feature representation level. In particular, devising privacy preservation testing methods to evaluate the success of adversarial identity obfuscation and related methods is a needed and not fully addressed research problem in cancer imaging and AI in healthcare at large.

Patient security. With the projected increase in clinical AI applications, adversarial learning based cybersecurity methodology becomes increasingly important to protect patients against the vulnerabilities inherent in clinically deployed deep learning solutions. Attacks can alter diagnostic markers on cancer imaging data, which can potentially result in diagnostic errors with dangerous consequences for the targeted patients. For instance, defences against adversarial examples (Liu et al., 2020b; Samangouei et al., 2018) or detection of imaging data that has been tampered with Mirsky et al. (2019) are areas where solutions based on adversarial methods will increasingly gain practical importance.

Model debiasing. The versatile ability of adversarial training to curate a model's latent space is likely to continue to increase in popularity due to the need to remove certain features in clinical AI models. For example, it is desirable to minimise a model's learned biases to increase the fairness of clinical models across patient populations (Lekadir et al., 2021). Such bias removal has been shown to be achievable via adversarial loss backpropagation (Zhang et al., 2018a; Li et al., 2021a). As Elazar and Goldberg (2018) point out, some residual biases may remain in a model's latent space after converged adversarial bias removal training. Therefore, research potential lies in automated test and evaluation methodology to assess the quantity of residual bias remaining in an adversarial networks after debiasing, particularly if applied to data unseen during training.

Generative model evaluation. A key aspect this survey observes is the absence of interpretable, standardised and exact evaluation methodology for synthetic data and generative models in the medical and cancer imaging domains. This is particularly noticeable for models without a narrow downstream task performance objective that can be used as surrogate evaluation metric nor a reconstruction objective that informs the evaluation technique. Generative models that generate a synthetic image with a clear reference value (i.e., a real image) can be evaluated based on the difference between reference and generated sample, e.g., via perceptual and reconstruction losses and metrics such as SSIM, PSNR, MSE, as discussed in Section 4.1. In the absence of such reference images, remaining methods at hand are image inspection techniques and real versus synthetic distribution comparisons, the latter including the Fréchet Inception Distance (FID) score (Heusel et al., 2017). The popularity of the FID metric for fidelity and diversity evaluation of synthetic data has largely translated from computer vision into medical imaging. The applicability of FID in the medical domain, nonetheless, is questionable, as it internally relies on an inception classifier pretrained on the ImageNet dataset consisting of 3-channel natural images as opposed to, for instance, grayscale images from radiological domains. This demonstrates a clear need for research on further evaluation methodologies of synthetic medical images. FID extensions that pre-train the internal classifier on medical imaging datasets are potential directions, but limited by the acquisition techniques, scope, modalities,

and, importantly, the size of these medical imaging datasets. Recent promising work proposed the automated generation of segmentation mask from GANs based on latent space exploration (Melas-Kyriazi et al., 2021). Such latent space inspection approaches can offer further potential for generative model evaluation, e.g., by helping to measure the number and difference between modes or by providing quality and diversity estimates of the segmentation masks (or other extractable pieces of information) that the model produces.

Patient treatment. Sections 4.3 and 4.4 have shown that adversarial models for cancer detection, classification, and localisation are, at least for particular organs and modalities, well explored research areas. These applications are mostly relevant in diagnostic activities, which comprise only one part of the clinical workflow. We encourage more research on GAN-based solutions in less explored subsequent clinical workflow steps such as oncological treatment planning and disease monitoring as elaborated in Section 4.5. For example, adversarial learning offers research potential in tumour profiling and intra- and inter-tumour heterogeneity assessment via anomaly detection within the latent space of adversarial models (Schlegl et al., 2019; Quiros et al., 2019). The high intra- and inter-tumour heterogeneity increases the difficulty of assessing and selecting targeted treatment options. Research potential exists in precisely encoding a tumour based on imaging and/or non-imaging patient and tumour data in an adversarial model's multi-dimensional latent space. For example, this can unlock vector search applications to find similarly encoded tumours in databases to inform on therapy selection, success probabilities, and progression patterns. Tumour progression modelling on image-level based on generative models such as GANs remains largely unexplored. Even though not strictly necessary (Xia et al., 2021), longitudinal and time-series cancer imaging datasets will likely trigger increased exploration of this research area once such data becomes available. For instance, given a tumour image at timepoint t_1 , a GAN can learn to simulate the tumour image at timepoint t_2 . To this end, generation of image-level counterfactuals (Pawlowski et al., 2020) as a clinically impactful solution for probing interventions. For instance, GANs can generate a tumour at t_2 given the tumour image at t_1 alongside multiple input conditions such as tumour growth rate, tumour type, and applied treatments.

6.7. Future perspectives and technology trends

6.7.1. Towards state-of-the-art GAN innovations in cancer imaging

In recent years, multiple novel adversarial networks have been introduced in the field of computer vision. A lesson learned from our survey is that many of these techniques are yet to be applied thoroughly to cancer imaging. These innovations open avenues in cancer imaging that extend upon the currently used methods shown in Fig. 4, for instance, enabling improved high-resolution image generation and input-conditioned image synthesis.

Overcoming dataset and computation limitations. For instance, the recent VQGAN (Esser et al., 2021) combines the efficiency of convolutional networks with the expressiveness of transformers, which model the composition of a reusable codebook of context-rich visual parts. This approach is particularly relevant to medical and cancer imaging, as it allows high-resolution image synthesis despite limited computing resources. Apart from containing high resolution images, cancer imaging datasets are often limited in the number of image, which may not suffice to train a GAN. In these cases, the potential issues are that during training there is convergence-failure, where the synthetic image quality is low and does not improve any further during training. While the adversarial loss often is non-interpretable not corresponding to synthetic image fidelity, diversity or condition adherence, also mode collapse may occur, where the generator has learned a particular mode to fool the discriminator instead of generating a high diversity of samples. These issues are not only a function of the GAN architecture and loss

function, but also of the size of the training dataset. FastGAN (Liu et al., 2020a) and SinGAN (Shaham et al., 2019) shows great promise to overcome this data scarcity cancer imaging problem. FastGAN (Liu et al., 2020a) uses self-supervised training of discriminator as encoder for regularisation and generates high-resolution images despite limited computing resources and dataset size. SinGAN (Shaham et al., 2019) generates multiple synthetic images based on only a single training image. This has wide applicability and can substantially increase the usefulness of even very small cancer imaging datasets via SinGAN-based data augmentation. A first successful applications of SinGAN and FastGAN to cancer imaging for polyp segmentation by Thambawita et al. (2022) shows the potential of using these models to generate not only a synthetic images, but also a corresponding segmentation mask by outputting an additional channel. This type of methodology enables training data generation for tumour detection, localisation and segmentation models without the need of conditioning the GAN on input segmentation masks.

Best practice combining GAN frameworks. As a vast amount of novel additions to the GANs framework has been suggested, some work (Brock et al., 2018) has focused on collecting the best working practices and combining them into novel architectures, which are promising and not yet widely applied to challenges in cancer imaging. For example, BigGAN (Brock et al., 2018) (a) scales model parameters by increasing the size of the feature maps, (b) applies large batch sizes, (c) uses self-attention based on SAGAN (Zhang et al., 2019a), (d) provides information about the class via class-conditional batch normalisation, and (e) uses hinge-loss. BigGAN and extensions thereof (e.g., Zhang et al. (2019b), Casanova et al. (2021) and Schonfeld et al. (2020)) achieve state-of-the-art performance on class-conditional image generation.

Extending on PGGAN (Karras et al., 2017) as shown in Fig. 4(m), another such example is StyleGAN (Karras et al., 2019) and its variants (Karras et al., 2021, 2020; Sauer et al., 2022), which accomplish state-of-the-art performance in conditional and unconditional computer vision image generation benchmarks. Yielding strong results, multiple architectural innovations have been introduced by the StyleGAN family, such as a style vector generating fully connected mapping network, adaptive instance normalisation, and, instead of sampling from a noise vector, moving the noise input to intermediate activation maps. These innovations can inform cancer image generation models and improve their latent space exploration capabilities e.g. allowing to compare different tumour types and manifestations.

Image-to-image translation. Image-to-image translation problems in cancer imaging are widely approached using commonly pix2pix (Isola et al., 2017) (paired) and cycleGAN (Zhu et al., 2017) (unpaired). Nonetheless, more recent models such as OASIS (Sushko et al., 2020), ResVit (Dalmaz et al., 2021), and StarGAN V2 (Choi et al., 2020) have been proposed, which are not only applicable to cancer imagery, but also have shown superior performance on computer vision benchmarks. ResVit (Dalmaz et al., 2021), for instance, diverges away from common CNN architectures with inductive biases by using a vision transformer architecture (Dosovitskiy et al., 2020) alongside an adversarial loss (Goodfellow et al., 2014) and the common L1 losses between source and target (Isola et al., 2017) and between source and reconstructed source (Zhu et al., 2017). StarGAN V2 (Choi et al., 2020) employs besides the adversarial and cycle consistency losses also a style reconstruction loss and a style diversification loss, while OASIS (Sushko et al., 2020) shows that a perceptual loss is not necessary given an adversarial loss and a segmentation-based discriminator.

6.7.2. GAN alternatives and complementary methods

Diffusion models. In image inpainting (Saharia et al., 2021a) and super resolution (Saharia et al., 2021b), the recently proposed and increasingly popular diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) have been shown to achieve state-of-the-art and competitive performances for computer vision benchmarks

and, thus, are an alternative to GANs. Diffusion models iteratively add noise to an image in a Markov chain of diffusion steps. Reversing this process, a noise vector z is gradually denoised and transformed into an image. While achieving promising generative modelling capabilities, it still takes longer to sample from diffusion models than from GANs due to multiple denoising steps, while also further work is needed to explore the interpretability of latent representations of diffusion models (Dhariwal and Nichol, 2021). A promising line of research suggests the combination of GANs with diffusion models to increase the stability and data efficiency of GAN training (Wang et al., 2022).

Variational autoencoders. GANs are commonly considered to achieve higher quality outputs than variational autoencoders (VAEs) (Kingma and Welling, 2013) at the cost of a training process more prone towards requiring manual intervention and tuning. A promising line of research improves upon vanilla VAE by exploring combinations of GANs and VAEs (Larsen et al., 2016; Makhzani et al., 2015). Extending on VAEs, Van Den Oord et al. (2017) proposed Vector Quantised Variational AutoEncoder (VQ-VAE), which learns discrete instead of continuous latent representations to avoid the issue of 'posterior collapse' that is common in VAEs. VQ-VAE has been shown to be an effective method for diverse high-quality synthetic image generation (Razavi et al., 2019). A promising extension combines VQ-VAE with transformers (Vaswani et al., 2017) for unsupervised anomaly detection and segmentation and demonstrates its potential for tumour segmentation in brain MRI (Pinaya et al., 2021).

Normalizing flows. The recently proposed Normalizing Flows (Rezende and Mohamed, 2015; Dinh et al., 2014, 2016) are an alternative deep generative model gaining increasing popularity for synthetic data generation tasks. As opposed to GANs and VAEs (implicit), Normalizing Flows explicitly learn the probability density function $p(x)$ and are trained via maximum likelihood estimation. Knowing $p(x)$, unobserved but realistic new data points can be sampled with exact likelihood estimates. Normalizing Flows have been shown to be combinable with GANs and the adversarial loss function, e.g., by being the building block of the generator network (Grover et al., 2018), and for image-to-image translation (Grover et al., 2020). To date, Normalizing Flows have seen less adoption in medical and cancer imaging than GANs, but promising initial applications exist. For example, Normalizing Flows have been proposed for uncertainty estimation of lung lesion segmentation (Selvan et al., 2020), counterfactual inference on brain MRI (Pawlowski et al., 2020), and low-dose CT image reconstruction (Denker et al., 2020).

Unsupervised domain adaptation. In unsupervised domain adaptation, self-training approaches are described as an alternative to domain adversarial losses. For example, state-of-the-art methods like HRDA (Hoyer et al., 2022b) and DaFormer (Hoyer et al., 2022a) show the effectiveness of self-training in domain-adaptive semantic segmentation. DaFormer uses a transformer encoder (Vaswani et al., 2017; Dosovitskiy et al., 2020) and transfers knowledge from source to target domain via a teacher network that generates pseudo-labels for the data from the target domain. A promising avenue of research combines self-training approaches and adversarial losses (Li et al., 2019c; Kim and Byun, 2020; Wang et al., 2020a).

Self-supervised learning. Given successes in learning useful representations from unlabelled data, self-supervised learning (SSL) approaches, such as BYOL (Grill et al., 2020), have become a common technique in the toolkit of deep learning researchers. Particularly when working with datasets limited in size or annotations, additional GAN-generated data can improve the learning of representations, upon which a downstream task model produces its predictions. SSL can provide an alternative, often computationally less expensive, means towards representation learning given a training task with objective function, where labels y and inputs x are extracted from an unlabelled dataset. A popular and powerful SSL method is contrastive learning, where a

model's latent space is learned by minimising the distance of similar samples and maximising the distance between dissimilar ones. Effective model pretraining methods such as SimCLR (Chen et al., 2020b) rely on such contrastive loss functions, which, e.g., maximise agreement between differently augmented views of the same image. Multiple recent studies propose the combination of GANs and self-supervised (Patel et al., 2021) and contrastive learning with promising results reporting improved performance and sample diversity, as well as reduced discriminator overfitting (Jeong and Shin, 2021; Kang and Park, 2020; Liu et al., 2021). In cancer imaging, for instance, this combination has been applied to address the problem of mode collapse while retaining phenotypic tumour features for the task of colour normalisation in histopathology images (Ke et al., 2021).

7. Conclusion

In closing, we emphasise the versatility and the resulting modality-independent wide applicability of the adversarial learning scheme of GANs. In this survey, we strive to consider and communicate this versatility by describing the wide variety of problems in the cancer imaging domain that can be approached with adversarial networks. For example, we highlight GAN and adversarial training solutions that range from unsupervised domain adaptation to patient privacy preserving distributed data synthesis, to adversarial segmentation mask discrimination, to multi-modal radiation dose estimation, amongst others.

Before reviewing and describing GAN and adversarial training solutions, we surveyed the literature to understand the current challenges in the field of cancer imaging with a focus on radiology, but without excluding non-radiology modalities common to cancer imaging. After screening and analysing the cancer imaging challenges, we grouped them into the challenge categories Data Scarcity and Usability, Data Access and Privacy, Data Annotation and Segmentation, Detection and Diagnosis, and Treatment and Monitoring. After categorisation, we surveyed the literature for adversarial networks applied to the field of cancer imaging and found 164 relevant publications, each of which we assigned to its respective cancer imaging challenge category. Finally, we provide a comprehensive analysis for each challenge and its assigned GAN-related publications to determine to what extent it has and can be solved using GANs and adversarial training. We further establish the *SynTRUST* framework for assessing the trustworthiness of medical image synthesis studies. Based on *SynTRUST*, we analyse 16 carefully selected cancer imaging challenge solutions. Notwithstanding the overall high level of rigour and validity of these studies, we are able to recommend a set of unaddressed trustworthiness improvements in order to guide future studies. To this end, we also highlight research potential for challenges where we were able to propose data synthesis or adversarial training solutions that have not yet been fully explored by the literature.

With our work, we strive to uncover and motivate promising lines of research in data synthesis and adversarial networks that we envision to ultimately benefit the field of cancer imaging in clinical practice.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952103.

References

- AAPM, 2016. 2016 NIH-AAPM-Mayo clinic low dose CT grand challenge. URL: <https://www.aapm.org/grandchallenge/lowdosect/>.
- Abdar, M., Pourpanah, F., Hussain, S., Rezaadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2020. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. arXiv preprint arXiv:2011.06225.
- Abhishek, K., Hamarneh, G., 2019. Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 71–80.
- Abramian, D., Eklund, A., 2019. Refacing: reconstructing anonymized facial features using GANs. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 1104–1108.
- Adamson, A.S., Smith, A., 2018. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 154 (11), 1247–1248.
- Addepalli, S., Nayak, G.K., Chakraborty, A., Radhakrishnan, V.B., 2020. DeGAN: Data-Enriching GAN for retrieving representative samples from a trained classifier. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 3130–3137.
- Ahmed, K.T., Sun, J., Yong, J., Zhang, W., 2021. Multi-omics data integration by generative adversarial network. *BioRxiv* <http://dx.doi.org/10.1101/2021.03.13.435251>.
- Almond, D., Chay, K.Y., Lee, D.S., 2004. The Costs of Low Birth Weight. Working Paper 10552, National Bureau of Economic Research, <http://dx.doi.org/10.3386/w10552>.
- Alshehri, R., Alshehri, A., 2021. Quantification of uncertainty in brain tumor segmentation using generative network and Bayesian active learning. In: VISIGRAPP (4: VISAPP). pp. 701–709.
- Alyafi, B., Diaz, O., Martí, R., 2020. DCGANs for realistic breast mass augmentation in x-ray mammography. In: Medical Imaging 2020: Computer-Aided Diagnosis, Vol. 11314. International Society for Optics and Photonics, 1131420.
- Angermann, Q., Bernal, J., Sánchez-Montes, C., Hammami, M., Fernández-Esparrach, G., Dray, X., Romain, O., Sánchez, F.J., Histace, A., 2017. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. Springer, pp. 29–41.
- Argenziano, G., Soyer, H., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M., et al., 2002. Dermoscopy: A Tutorial, Vol. 16. EDRA, Medical Publishing & New Media.
- Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 214–223.
- Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: Medical image translation using GANs. *Comput. Med. Imaging Graph.* 79, 101684.
- Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* 38 (2), 915–931.
- Arnold, M., Rutherford, M.J., Bardot, A., Ferlay, J., Andersson, T.M., Myklebust, T.Å., Tervonen, H., Thursfield, V., Ransom, D., Shack, L., et al., 2019. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol.* 20 (11), 1493–1505.
- Arora, S., Risteski, A., Zhang, Y., 2018. Do GANs learn the distribution? some theory and empirics. In: International Conference on Learning Representations.
- Babier, A., Boutilier, J.J., Sharpe, M.B., McNiven, A.L., Chan, T.C., 2018. Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms. *Phys. Med. Biol.* 63 (10), 105004.
- Bae, H., Jung, D., Yoon, S., 2020. AnomIGAN: Generative adversarial networks for anonymizing private medical data. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, Vol. 25. World Scientific, pp. 563–574.
- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4 (1), 1–13.
- Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
- Balagurunathan, Y., Kumar, V., Gu, Y., Kim, J., Wang, H., Liu, Y., Goldgof, D.B., Hall, L.O., Korn, R., Zhao, B., et al., 2014. Test-retest reproducibility analysis of lung CT image features. *J. Digit. Imaging* 27 (6), 805–823.
- Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J., 2013. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis. Springer, pp. 63–86.
- Barbaro, B., Leccisotti, L., Vecchio, F.M., Di Matteo, M., Serra, T., Salsano, M., Poscia, A., Coco, C., Persiani, R., Alfieri, S., et al., 2017. The potential predictive value of MRI and PET-CT in mucinous and nonmucinous rectal cancer to identify patients at high risk of metastatic disease. *Br. J. Radiol.* 90 (1069), 20150836.
- Baur, C., Albarqouni, S., Navab, N., 2018. MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv preprint arXiv:1804.04338.

- Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., Greene, C.S., 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ.: Cardiovasc. Qual. Outcomes* 12 (7), e005122.
- Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M.J., West, R.B., van de Rijn, M., Koller, D., 2011. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* 3 (108), 108ra113.
- Becker, A.S., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., Konukoglu, E., 2019. Injecting and removing suspicious in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images. *Eur. J. Radiol.* 120, 108649.
- Ben-Cohen, A., Klang, E., Raskin, S.P., Amitai, M.M., Greenspan, H., 2017. Virtual PET images from CT data using deep convolutional networks: initial results. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, pp. 49–57.
- Benhammou, Y., Achhab, B., Herrera, F., Tabik, S., 2020. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing* 375, 9–24.
- Benson, S., Beets-Tan, R., 2020. GAN-based anomaly detection in multi-modal MRI images. *BioRxiv* <http://dx.doi.org/10.1101/2020.07.10.197087>.
- Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* 43, 99–111.
- Bernard, O., Lalonde, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imaging* 37 (11), 2514–2525.
- Beutel, A., Chen, J., Zhao, Z., Chi, E.H., 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.
- Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al., 2019. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: Cancer J. Clin.* 69 (2), 127–157.
- Bi, L., Kim, J., Kumar, A., Feng, D., Fulham, M., 2017. Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs). In: *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Springer, pp. 43–51.
- Bica, I., Jordon, J., van der Schaar, M., 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*.
- Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.-W., Han, X., Heng, P.-A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*.
- Billot, B., Robinson, E., Dalca, A.V., Iglesias, J.E., 2020. Partial volume segmentation of brain MRI scans of any resolution and contrast. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 177–187.
- Bischoff-Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema-Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M., et al., 2007. A technique for the deidentification of structural brain MR images. *Hum. Brain Mapp.* 28 (9), 892–903.
- Bissoto, A., Perez, F., Valle, E., Avila, S., 2018. Skin lesion synthesis with generative adversarial networks. In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, pp. 294–302.
- Blake, C., 1998. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al., 2020. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* 7 (1), 1–14.
- Borji, A., 2019. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.* 179, 41–65.
- Borji, A., 2021. Pros and cons of GAN evaluation measures: New developments. *arXiv preprint arXiv:2103.09396*.
- Brady, A.P., 2017. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging* 8 (1), 171–182.
- van den Brandt, P.A., Goldbohm, R.A., Veer, P.V., Volovics, A., Hermus, R.J., Sturmans, F., 1990. A large-scale prospective cohort study on diet and cancer in The Netherlands. *J. Clin. Epidemiol.* 43 (3), 285–295.
- Brennan, P., Silman, A., 1992. Statistical methods for assessing observer variability in clinical measures. *BMJ: Br. Med. J.* 304 (6840), 1491.
- Brock, A., Donahue, J., Simonyan, K., 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.
- Bromley, J., Guyon, I., LeCun, Y., 1994. Signature verification using a “siamese” time delay neural network.
- Bu, T., Yang, Z., Jiang, S., Zhang, G., Zhang, H., Wei, L., 2020. 3D conditional generative adversarial network-based synthetic medical image augmentation for lung nodule detection. *Int. J. Imaging Syst. Technol.*
- Buda, M., Saha, A., Walsh, R., Gbate, S., Li, N., Swiecicki, A., Lo, J.Y., Mazurowski, M.A., 2021. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Netw. Open* 4 (8), e2119100.
- Caballo, M., Pangallo, D.R., Mann, R.M., Sechopoulos, I., 2020. Deep learning-based segmentation of breast masses in dedicated breast CT imaging: radiomic feature stability between radiologists and artificial intelligence. *Comput. Biol. Med.* 118, 103629.
- Casanova, A., Careil, M., Verbeek, J., Drozdal, M., Romero Soriano, A., 2021. Instance-conditioned gan. *Adv. Neural Inf. Process. Syst.* 34, 27517–27529.
- Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. *Nature Commun.* 11 (1), 1–10.
- Cem Birbiri, U., Hamidinekoo, A., Grall, A., Malcolm, P., Zwiggelaar, R., 2020. Investigating the performance of generative adversarial networks for prostate tissue detection and segmentation. *J. Imaging* 6 (9), 83.
- Chaitanya, K., Karani, N., Baumgartner, C.F., Erdil, E., Becker, A., Donati, O., Konukoglu, E., 2021. Semi-supervised task-driven data augmentation for medical image segmentation. *Med. Image Anal.* 68, 101934.
- Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G., 2020. Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. *arXiv preprint arXiv:2012.00926*.
- Chang, Q., Qu, H., Zhang, Y., Sabuncu, M., Chen, C., Zhang, T., Metaxas, D.N., 2020a. Synthetic learning: Learn from distributed asynchronized discriminator GAN without sharing medical image data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13856–13866.
- Chang, Q., Yan, Z., Baskaran, L., Qu, H., Zhang, Y., Zhang, T., Zhang, S., Metaxas, D.N., 2020b. Multi-modal AsynDGAN: Learn from distributed medical image data without sharing private information. *arXiv preprint arXiv:2012.08604*.
- Chaudhari, P., Agrawal, H., Kotecha, K., 2019. Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Comput.* 1–11.
- Chen, S., Jia, R., Qi, G.-J., 2020a. Improved techniques for model inversion attack. *arXiv preprint arXiv:2010.04092*.
- Chen, J., Konrad, J., Ishwar, P., 2018a. Vgan-based image representation learning for privacy-preserving facial expression recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 1570–1579.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*. PMLR, pp. 1597–1607.
- Chen, X., Li, Y., Yao, L., Adeli, E., Zhang, Y., 2021. Generative adversarial U-net for domain-free medical image augmentation. *arXiv preprint arXiv:2101.04793*.
- Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., Konukoglu, E., 2018b. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint: arXiv:1806.05452*.
- Chen, L., Song, H., Wang, C., Cui, Y., Yang, J., Hu, X., Zhang, L., 2019. Liver tumor segmentation in CT volumes using an adversarial densely connected network. *BMC Bioinformatics* 20 (16), 1–13.
- Chen, L., Yang, Q., Bao, J., Liu, D., Huang, X., Wang, J., 2017. Direct comparison of PET/CT and MRI to predict the pathological response to neoadjuvant chemotherapy in breast cancer: a meta-analysis. *Sci. Rep.* 7 (1), 1–10.
- Chi, Y., Bi, L., Kim, J., Feng, D., Kumar, A., 2018. Controlled synthesis of dermoscopic images via a new color labeled generative style transfer network to enhance melanoma segmentation. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE*, pp. 2591–2594.
- Choi, Y., Uh, Y., Yoo, J., Ha, J.-W., 2020. Stargan v2: Diverse image synthesis for multiple domains. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8188–8197.
- Chong, M.J., Forsyth, D., 2020. Effectively unbiased FID and inception score and where to find them. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6070–6079.
- Choyke, P., Turkbey, B., Pinto, P., Merino, M., Wood, B., 2016. Data from PROSTATE-MRI. *Cancer Imaging Arch.* 9, Available Online: <http://doi.org/10.7937/K9/TCIA.2016.6046GUDv>.
- Chuquicusma, M.J., Hussein, S., Burt, J., Bagci, U., 2018. How to fool radiologists with generative adversarial networks? a visual Turing test for lung cancer diagnosis. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 240–244.
- GireAan, D., Meier, U., Masci, J., Schmidhuber, J., 2012. Multi-column deep neural network for traffic sign classification. *Neural Netw.* 32, 333–338.
- Giresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 411–418.
- Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al., 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163 (2), 506–519.
- Cirillo, M.D., Abramian, D., Eklund, A., 2020. Vox2Vox: 3D-GAN for brain tumour segmentation. *arXiv preprint arXiv:2003.13653*.

- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26 (6), 1045–1057.
- Coccosco, C.A., Kollokian, V., Kwan, R.K.-S., Pike, G.B., Evans, A.C., 1997. Brainweb: Online interface to a 3D MRI simulated brain database. In: *NeuroImage*. Citeseer.
- Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kallou, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 168–172.
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kallou, A., Liopyris, K., Marchetti, M., et al., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*.
- Cohen, J.P., Luck, M., Honari, S., 2018a. Distribution matching losses can hallucinate features in medical image translation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 529–536.
- Cohen, J.P., Luck, M., Honari, S., 2018b. How to cure cancer (in images) with unpaired image translation. In: *International Conference on Medical Imaging with Deep Learning (MIDL 2018)–Abstract Track*.
- Combialia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.
- Creswell, A., Pouplin, A., Bharath, A.A., 2018. Denoising adversarial autoencoders: classifying skin lesions using limited labelled training data. *IET Comput. Vis.* 12 (8), 1105–1111.
- Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shomovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P., 2014. The reactome pathway knowledgebase. *Nucl. Acids Res.* 42 (Database issue), D472–477.
- Cuocolo, R., Caruso, M., Perillo, T., Uggla, L., Petretta, M., 2020. Machine learning in oncology: A clinical appraisal. *Cancer Lett.* 481, 55–62.
- Dalmaz, O., Yurt, M., Çukur, T., 2021. Resvit: Residual vision transformers for multi-modal medical image synthesis. *arXiv preprint arXiv:2106.16031*.
- Dashtbani, M., Li, W., 2020. Predicting risk of hospital readmission for comorbidity patients through a novel deep learning framework. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
- Denker, A., Schmidt, M., Leuschner, J., Maass, P., Behrmann, J., 2020. Conditional normalizing flows for low-dose computed tomography image reconstruction. *arXiv preprint arXiv:2006.06270*.
- Desai, S.D., Giraddi, S., Verma, N., Gupta, P., Ramya, S., 2020. Breast cancer detection using GAN for limited labeled dataset. In: 2020 12th International Conference on Computational Intelligence and Communication Networks. CICN, IEEE, pp. 34–39.
- DeVries, T., Romero, A., Pineda, L., Taylor, G.W., Drozdal, M., 2019. On the evaluation of conditional GANs. *arXiv preprint arXiv:1907.08175*.
- Dhariwal, P., Nichol, A., 2021. Diffusion models beat GANs on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., pp. 8780–8794. URL: <https://proceedings.neurips.cc/paper/2021/file/49ad23d1ce9fab4bd8d77d02681df5cfa-Paper.pdf>.
- Diaz, O., Kushibar, K., Osuala, R., Linares, O., Garrucho, L., Igual, L., Radeva, P., Prior, F., Gkontra, P., Lekadir, K., 2021. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Phys. Medica* 83, 25–37. <http://dx.doi.org/10.1016/j.ejmp.2021.02.007>.
- Dimitriou, N., Arandjelović, O., Harrison, D.J., Caie, P.D., 2018. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *NPJ Digit. Med.* 1 (1), 1–9.
- Dinh, L., Krueger, D., Bengio, Y., 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.
- Dinh, L., Sohl-Dickstein, J., Bengio, S., 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.
- Doman, K., Konishi, T., Mekada, Y., 2020. Lesion image synthesis using DCGANs for metastatic liver cancer detection. *Deep Learn. Med. Image Anal.* 95–106.
- Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W.J., Liu, T., Yang, X., 2019. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Med. Phys.* 46 (5), 2157–2168.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Drew, T., Vö, M.L.-H., Wolfe, J.M., 2013. The invisible gorilla strikes again: Sustained inattention blindness in expert observers. *Psychol. Sci.* 24 (9), 1848–1853.
- Durán, J.M., Jongsma, K.R., 2021. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *J. Med. Ethics* 47 (5), 329–335.
- Dwork, C., 2006. Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (Eds.), *Automata, Languages and Programming*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–12.
- Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1), 207–210.
- Edupuganti, V., Mardani, M., Cheng, J., Vasanawala, S., Pauly, J., 2019. Uncertainty analysis of VAE-GANs for compressive medical imaging. *arXiv preprint arXiv:1901.11228*.
- Elazab, A., Wang, C., Gardezi, S.J.S., Bai, H., Hu, Q., Wang, T., Chang, C., Lei, B., 2020. GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR images. *Neural Netw.* 132, 321–332. <http://dx.doi.org/10.1016/j.neunet.2020.09.004>.
- Elazar, Y., Goldberg, Y., 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.
- Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H., Feinstein, A.R., 1994. Variability in radiologists' interpretations of mammograms. *N. Engl. J. Med.* 331 (22), 1493–1499.
- Esser, P., Rombach, R., Ommer, B., 2021. Taming transformers for high-resolution image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12873–12883.
- Esteve, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. *Nat. Med.* 25 (1), 24–29.
- Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., Hermjakob, H., 2017. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18 (1), 142.
- Farnia, F., Ozdaglar, A., 2020. GANs may have no nash equilibria. *arXiv preprint arXiv:2002.09124*.
- Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F., 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136 (5), E359–E386.
- Fernandes, K., Cardoso, J.S., Fernandes, J., 2017. Transfer learning with partial observability applied to cervical cancer screening. In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, pp. 243–250.
- Fischer, B.M., Olsen, M.W., Ley, C.D., Klausen, T.L., Mortensen, J., Højgaard, L., Kristjansen, P.E., 2006. How few cancer cells can be detected by positron emission tomography? a frequent question addressed by an in vitro study. *Eur. J. Nucl. Med. Mol. Imaging* 33 (6), 697–702.
- Fitzpatrick, J.M., 1998. RIRE - retrospective image registration evaluation. URL: <https://www.insight-journal.org/rire/>.
- Foroozandeh, M., Eklund, A., 2020. Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE. *arXiv preprint arXiv:2009.05946*.
- Fossen-Romsaas, S., Storm-Johannessen, A., Lundervold, A.S., 2020. Synthesizing skin lesion images using CycleGANs—a case study.
- Frangioni, J.V., 2008. New technologies for human cancer imaging. *J. Clin. Oncol.* 26 (24), 4012.
- Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321, 321–331.
- Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* 36 (4), 193–202. <http://dx.doi.org/10.1007/BF00344251>.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International Conference on Machine Learning*. PMLR, pp. 1050–1059.
- GANIN, Y., LEMPITSKY, V., 2015. Unsupervised domain adaptation by backpropagation. In: *International Conference on Machine Learning*. PMLR, pp. 1180–1189.
- Gao, C., Clark, S., Furst, J., Raicu, D., 2019. Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection. In: *Medical Imaging 2019: Computer-Aided Diagnosis*, Vol. 10950. International Society for Optics and Photonics, 109501K.
- GDPR, 2016. European Parliament and Council of European Union (2016) regulation (EU) 2016/679. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679>. [Online; accessed 26-November-2020].
- Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., Xiong, M., 2020. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Front. Genet.* 11, 1578. <http://dx.doi.org/10.3389/fgene.2020.585804>.
- Ghorbani, A., Natarajan, V., Coz, D., Liu, Y., 2020. DermGAN: synthetic generation of clinical skin images with pathology. In: *Machine Learning for Health Workshop*. PMLR, pp. 155–170.
- Ghosal, S.S., Sarkar, I., El Hallaoui, I., 2020. Lung nodule classification using convolutional autoencoder and clustering augmented learning method (CALM). In: *HSDM@ WSDM*. pp. 19–26.
- Giacomello, E., Loiacono, D., Mainardi, L., 2020. Brain MRI tumor segmentation with adversarial networks. In: 2020 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8.
- Gilles, F.H., Tavaré, C.J., Laurence, E.B., Burger, P.C., Yates, A.J., Pollack, I.F., Finlay, J.L., 2008. Pathologist interobserver variability of histologic features in childhood brain tumors: results from the CCG-945 study. *Pediatr. Dev. Pathol.* 11 (2), 108–117.

- Gohagan, J.K., Marcus, P.M., Fagerstrom, R.M., Pinsky, P.F., Kramer, B.S., Prorok, P.C., Ascher, S., Bailey, W., Brewer, B., Church, T., et al., 2005. Final results of the lung screening study, a randomized feasibility study of spiral CT versus chest X-ray screening for lung cancer. *Lung Cancer* 47 (1), 9–15.
- Gohagan, J., Marcus, P., Fagerstrom, R., Pinsky, P., Kramer, B., Prorok, P., Group, L.S.R.R., et al., 2004. Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs chest radiograph: the Lung Screening Study of the National Cancer Institute. *Chest* 126 (1), 114–121.
- Goldsborough, P., Pawlowski, N., Caicedo, J., Singh, S., Carpenter, A., 2017. CytoGAN: Generative modeling of cell images. <http://dx.doi.org/10.1101/227645>.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*, Vol. 1. MIT Press, <http://www.deeplearningbook.org>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680.
- Grall, A., Hamidineko, A., Malcolm, P., Zwiggelaar, R., 2019. Using a conditional generative adversarial network (cGAN) for prostate segmentation. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer, pp. 15–25.
- Grill, J.-B., Strub, F., Althé, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent: a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* 33, 21271–21284.
- Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M., 2016. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375 (12), 1109–1112.
- Grover, A., Chute, C., Shu, R., Cao, Z., Ermon, S., 2020. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. pp. 4028–4035.
- Grover, A., Dhar, M., Ermon, S., 2018. Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- Gu, Y., Zeng, Z., Chen, H., Wei, J., Zhang, Y., Chen, B., Li, Y., Qin, Y., Xie, Q., Jiang, Z., et al., 2020. MedSRGAN: medical images super-resolution using generative adversarial networks. *Multimedia Tools Appl.* 79, 21815–21840.
- Guan, S., Loew, M., 2019. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. *J. Med. Imaging* 6 (3), 031411.
- Guerraoui, R., Guirguis, A., Kermarrec, A.-M., Merrer, E.L., 2020. FeGAN: Scaling distributed GANs. In: *Proceedings of the 21st International Middleware Conference*. pp. 193–206.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein GANs. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dcd52936e27cbd0ff683d6-Paper.pdf>.
- Hadjiliski, L., Cho, H.-c., Chan, H.-P., Sahiner, B., Helvie, M.A., Paramagul, C., Nees, A.V., 2012. Inter-and intra-observer variability of radiologists evaluating CBIR systems. In: *International Workshop on Digital Mammography*. Springer, pp. 482–489.
- Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAviney, R., Young, K.C., 2020. OTMAM mammography image database: A large-scale resource of mammography images and clinical data. *Radiol. Artif. Intell.* e200103.
- Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., Nakayama, H., 2018. GAN-based synthetic brain MR image generation. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 734–738.
- Han, C., Kitamura, Y., Kudo, A., Ichinose, A., Rundo, L., Furukawa, Y., Umemoto, K., Li, Y., Nakayama, H., 2019a. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. [arXiv:1906.04962](https://arxiv.org/abs/1906.04962).
- Han, C., Murao, K., Satoh, S., 2019b. Learning more with less: GAN-based medical image augmentation. [arXiv:1904.00838](https://arxiv.org/abs/1904.00838).
- Han, X., Qi, L., Yu, Q., Zhou, Z., Zheng, Y., Shi, Y., Gao, Y., 2021. Deep symmetric adaptation network for cross-modal medical image segmentation. [arXiv preprint arXiv:2101.06853](https://arxiv.org/abs/2101.06853).
- Han, C., Rundo, L., Araki, R., Furukawa, Y., Mauri, G., Nakayama, H., Hayashi, H., 2020. Infinite brain MR images: PGGAN-based data augmentation for tumor detection. In: *Neural Approaches to Dynamics of Signal Exchanges*. Springer, pp. 291–303.
- Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., Hayashi, H., 2019c. Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. *IEEE Access* 7, 156966–156977.
- Hanahan, D., Weinberg, R.A., 2000. The hallmarks of cancer. *Cell* 100 (1), 57–70.
- Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell* 144 (5), 646–674.
- Hardy, C., Le Merrer, E., Sericola, B., 2019. Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. In: *2019 IEEE International Parallel and Distributed Processing Symposium. IPDPS, IEEE*, pp. 866–877.
- Hasani, N., Morris, M.A., Rhamim, A., Summers, R.M., Jones, E., Siegel, E., Saboury, B., 2022. Trustworthy artificial intelligence in medical imaging. *PET Clin.* 17 (1), 1–12.
- Haubold, P., Hosch, R., Umuthu, L., Wetter, A., Haubold, P., Radbruch, A., Forsting, M., Nensa, F., Koitka, S., 2021. Contrast agent dose reduction in computed tomography with deep learning using a conditional generative adversarial network. *Eur. Radiol.* 1–9.
- He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y., 2019. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. pp. 8417–8424.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P., 2001. The digital database for screening mammography, IWDM-2000. In: *Fifth International Workshop on Digital Mammography*. Medical Physics Publishing, ISBN: 1-930524-00-5, pp. 212–218.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. [arXiv preprint arXiv:1706.08500](https://arxiv.org/abs/1706.08500).
- HIPAA, 1996. The health insurance portability and accountability act of 1996. *Public Law* 104, 191, <https://www.govinfo.gov/content/pkg/PLAW-104publ191/html/PLAW-104publ191.htm>.
- Hirano, H., Minagi, A., Takemoto, K., 2021. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* 21 (1), 1–13.
- Hitaj, B., Ateniese, G., Perez-Cruz, F., 2017. Deep models under the GAN: information leakage from collaborative deep learning. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. pp. 603–618.
- Hjelm, R.D., Jacob, A.P., Che, T., Trischler, A., Cho, K., Bengio, Y., 2017. Boundary-seeking generative adversarial networks. [arXiv preprint arXiv:1702.08431](https://arxiv.org/abs/1702.08431).
- Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33, 6840–6851.
- Hognon, C., Tixier, F., Gallinato, O., Colin, T., Visvikis, D., Jaouen, V., 2019. Standardization of multicentric image datasets with generative adversarial networks. In: *IEEE Nuclear Science Symposium and Medical Imaging Conference 2019*.
- Hopper, K.D., Kasales, C., Van Slyke, M.A., Schwartz, T.A., TenHave, T.R., Jozefiak, J.A., 1996. Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR. Am. J. Roentgenol.* 167 (4), 851–854.
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. *Nat. Rev. Cancer* 18 (8), 500–510.
- Hoyer, L., Dai, D., Van Gool, L., 2022a. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9924–9935.
- Hoyer, L., Dai, D., Van Gool, L., 2022b. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. [arXiv preprint arXiv:2204.13132](https://arxiv.org/abs/2204.13132).
- Hu, X., Chung, A.G., Fieguth, P., Khalvati, F., Haider, M.A., Wong, A., 2018b. Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. [arXiv preprint arXiv:1811.05817](https://arxiv.org/abs/1811.05817).
- Hu, X., Guo, R., Chen, J., Li, H., Waldmannstetter, D., Zhao, Y., Li, B., Shi, K., Menze, B., 2020. Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images. *IEEE J. Biomed. Health Inf.* 24 (9), 2599–2608.
- Hu, J., Shen, L., Sun, G., 2018a. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7132–7141.
- Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: *Proceedings of the European Conference on Computer Vision. ECCV*, pp. 172–189.
- Huang, S., Yang, J., Fong, S., Zhao, Q., 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* 471, 61–71.
- Hukkelås, H., Mester, R., Lindseth, F., 2019. Deepprivacy: A generative adversarial network for face anonymization. In: *International Symposium on Visual Computing*. Springer, pp. 565–578.
- Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.-Y., Yang, M.H., 2019. Adversarial learning for semi-supervised semantic segmentation. In: *29th British Machine Vision Conference, BMVC 2018*.
- Huo, Y., Xu, Z., Bao, S., Bermudez, C., Moon, H., Parvathaneni, P., Moyo, T.K., Savona, M.R., Assad, A., Abramson, R.G., et al., 2018. Splenomegaly segmentation on multi-modal MRI using deep convolutional networks. *IEEE Trans. Med. Imaging* 38 (5), 1185–1196.
- Huq, A., Pervin, M.T., 2020. Analysis of adversarial attacks on skin cancer recognition. In: *2020 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, pp. 1–4.
- Huynh, E., Hosny, A., Guthrie, C., Bitterman, D.S., Petit, S.F., Haas-Kogan, D.A., Kann, B., Aerts, H.J., Mak, R.H., 2020. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* 17 (12), 771–781.
- Hwang, U., Choi, S., Lee, H.-B., Yoon, S., 2017. Adversarial training for disease prediction from electronic health records with missing data. [arXiv preprint arXiv:1711.04126](https://arxiv.org/abs/1711.04126).

- Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.
- Itri, J.N., Tappouni, R.R., McEachern, R.O., Pesch, A.J., Patel, S.H., 2018. Fundamentals of diagnostic error in imaging. *Radiographics* 38 (6), 1845–1865.
- IXI Dataset, 2007. IXI Dataset by brain-development.org. URL: <http://brain-development.org/ixi-dataset/>.
- Jaeger, S., Karagyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al., 2013. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* 33 (2), 233–245.
- JAMIT Japanese Society of Medical Imaging Technology, JAMIT CAD Contest. URL: <http://www.jamit.jp/meetinginfo/cad.html>.
- Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. *CA: Cancer J. Clin.* 61 (2), 69–90.
- Jendele, L., Skopek, O., Becker, A.S., Konukoglu, E., 2019. Adversarial augmentation for enhancing classification of mammography images. *arXiv preprint arXiv:1902.07762*.
- Jeong, J., Shin, J., 2021. Training GANs with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*.
- Jiang, Y., Chang, S., Wang, Z., 2021. TransGAN: Two transformers can make one strong GAN. *arXiv:2102.07074*.
- Jiang, J., Hu, Y.-C., Tyagi, N., Zhang, P., Rimmer, A., Mageras, G.S., Deasy, J.O., Veeraraghavan, H., 2018. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 777–785.
- Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J., 2018. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 732–740.
- Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3 (1), 1–9.
- Jolicœur-Martineau, A., 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*.
- Jordon, J., Yoon, J., Van Der Schaar, M., 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations.
- Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., Zhavoronkov, A., 2017a. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8 (7), 10883–10890.
- Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., Zhavoronkov, A., 2017b. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* 14 (9), 3098–3104.
- Kaiser, B., Albarqouni, S., 2019. MRI to CT translation with GANs. *arXiv preprint arXiv:1901.05259*.
- Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2 (6), 305–311.
- Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.
- Kanayama, T., Kurose, Y., Tanaka, K., Aida, K., Satoh, S., Kitsuregawa, M., Harada, T., 2019. Gastric cancer detection from endoscopic images using synthesis by GAN. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 530–538.
- Kang, M., Park, J., 2020. Contragan: Contrastive learning for conditional image generation. *Adv. Neural Inf. Process. Syst.* 33, 21357–21369.
- Kang, D., Park, J.E., Kim, Y.-H., Kim, J.H., Oh, J.Y., Kim, J., Kim, Y., Kim, S.T., Kim, H.S., 2018. Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. *Neuro-Oncology* 20 (9), 1251–1261.
- Kansal, S., Goel, S., Bhattacharya, J., Srivastava, V., 2020. Generative adversarial network-convolution neural network based breast cancer classification using optical coherence tomographic images. *Laser Phys.* 30 (11), 115601.
- Kapil, A., Meier, A., Zuraw, A., Steele, K.E., Rebelatto, M.C., Schmidt, G., Briue, N., 2018. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. *Sci. Rep.* 8 (1), 1–10.
- Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 34, 852–863.
- Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020. Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119.
- Kather, J.N., Halama, N., Marx, A., 2018. 100,000 Histological images of human colorectal cancer and healthy tissue. <http://dx.doi.org/10.5281/zenodo.1214456>.
- Kazemifar, S., Barragán Montero, A.M., Souris, K., Rivas, S.T., Timmerman, R., Park, Y.K., Jiang, S., Geets, X., Sterpin, E., Owrangi, A., 2020. Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors. *J. Appl. Clin. Med. Phys.* 21 (5), 76–86.
- Kazemina, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2020. GANs for medical image analysis. *Artif. Intell. Med.* 101938.
- Kazuhiro, K., Werner, R.A., Toriumi, F., Javadi, M.S., Pomper, M.G., Solnes, L.B., Verde, F., Higuchi, T., Rowe, S.P., 2018. Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. *Tomography* 4 (4), 159.
- Ke, J., Shen, Y., Liang, X., Shen, D., 2021. Contrastive learning based stain normalization across multiple tumor in histopathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 571–580.
- Kearney, V., Chan, J.W., Wang, T., Perry, A., Descovich, M., Morin, O., Yom, S.S., Solberg, T.D., 2020a. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Sci. Rep.* 10 (1), 1–8.
- Kearney, V., Ziemer, B.P., Perry, A., Wang, T., Chan, J.W., Ma, L., Morin, O., Yom, S.S., Solberg, T.D., 2020b. Attention-aware discrimination for MR-to-CT image translation using cycle-consistent generative adversarial networks. *Radiol. Artif. Intell.* 2 (2), e190027.
- Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17 (1), 1–9.
- Kim, M., Byun, H., 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12975–12984.
- Kim, K.H., Do, W.-J., Park, S.-H., 2018a. Improving resolution of MR images with an adversarial network incorporating images with different contrast. *Med. Phys.* 45 (7), 3120–3131.
- Kim, B.N., Dolz, J., Jodoin, P.-M., Desrosiers, C., 2019a. Privacy-net: An adversarial approach for identity-obfuscated segmentation of medical images. *arXiv preprint arXiv:1909.04087*.
- Kim, H.S., Jung, J., Jeong, C., Kwak, J., Choi, E., Lee, S., Yoon, S., Cho, B., 2019b. Prediction of hepatic parenchymal change in gd-EOB-DTPA MR images after stereotactic body radiation therapy by cycle GAN deep neural network. *Int. J. Radiat. Oncol. Biol. Phys.* 105, E135. <http://dx.doi.org/10.1016/j.ijrobp.2019.06.2171>.
- Kim, S., Kim, B., Park, H., 2020. Synthesis of brain tumor MR images for learning data augmentation. *arXiv preprint arXiv:2003.07526*.
- Kim, R., Ock, C.-Y., Keam, B., Kim, T.M., Kim, J.H., Paeng, J.C., Kwon, S.K., Hah, J.H., Kwon, T.-K., Kim, D.-W., et al., 2016. Predictive and prognostic value of PET/CT imaging post-chemoradiotherapy and clinical decision-making consequences in locally advanced head & neck squamous cell carcinoma: a retrospective study. *BMC Cancer* 16 (1), 1–9.
- Kim, M., Oh, I., Ahn, J., 2018b. An improved method for prediction of cancer prognosis by network learning. *Genes* 9, 478. <http://dx.doi.org/10.3390/genes9100478>.
- Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Klaver, C.E., Bulkman, N., Drillenburger, P., Grabsch, H.I., van Grieken, N.C., Karrenbeld, A., Koens, L., van Lijnschoten, I., Meijer, J., Nagtegaal, I.D., et al., 2020. Interobserver, intraobserver, and interlaboratory variability in reporting pT4a colon cancer. *Virchows Archiv* 476 (2), 219–230.
- Kodali, N., Abernethy, J., Hays, J., Kira, Z., 2017. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*.
- Kohl, S., Bonekamp, D., Schlemmer, H.-P., Yaqubi, K., Hohenfellner, M., Hadaschik, B., Radtke, J.-P., Maier-Hein, K., 2017. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*.
- Koike, Y., Anetai, Y., Takegawa, H., Ohira, S., Nakamura, S., Tanigawa, N., 2020. Deep learning-based metal artifact reduction using cycle-consistent adversarial network for intensity-modulated head and neck radiation therapy treatment planning. *Phys. Medica* 78, 8–14.
- Korkinof, D., Harvey, H., Heindl, A., Karpati, E., Williams, G., Rijken, T., Kecske, P., Glocker, B., 2020. Perceived realism of high resolution generative adversarial network derived synthetic mammograms. *Radiol. Artif. Intell.* e190181.
- Kornblau, S.M., Tibes, R., Qiu, Y.H., Chen, W., Kantarjian, H.M., Andreeff, M., Coombes, K.R., Mills, G.B., 2009. Functional proteomic profiling of AML predicts response and survival. *Blood* 113 (1), 154–164.
- Korpihakola, J., Sipilä, T., Puuska, S., Kokkonen, T., 2020. One-pixel attack deceives automatic detection of breast cancer. *arXiv preprint arXiv:2012.00517*.
- Krause, J., Grabsch, H.I., Kloor, M., Jendrusch, M., Echle, A., Buelow, R.D., Boor, P., Luedde, T., Brinker, T.J., Trautwein, C., et al., 2021. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *J. Pathol.* 254 (1), 70–79.

- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.
- Kuang, Y., Lan, T., Peng, X., Selasi, G.E., Liu, Q., Zhang, J., 2020. Kuang. *IEEE Access* 8, 77725–77734.
- Kuijff, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseht, M., Belyaev, M., Cardoso, M.J., Casamijana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Trans. Med. Imaging* 38 (11), 2556–2568.
- Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imaging* 36 (7), 1550–1560.
- Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Veta, M., 2019. Learning domain-invariant representations of histological images. *Front. Med.* 6, 162.
- Lakshminarayanan, B., Pritzel, A., Blundell, C., 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.
- Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 48 (4), 441–446.
- LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., et al., 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv* <http://dx.doi.org/10.1101/2019.12.13.19014902>.
- Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., Chen, Y., Zhou, X., 2020. Generative adversarial networks and its applications in biomedical informatics. *Front. Public Health* 8, 164.
- Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E., 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci.* 117 (23), 12592–12594.
- Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O., 2016. Autoencoding beyond pixels using a learned similarity metric. In: *International Conference on Machine Learning*. PMLR, pp. 1558–1566.
- LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1 (4), 541–551.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324.
- Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4681–4690.
- Lee, R.S., Gimenez, F., Hoogi, A., Rubin, D., 2016. Curated breast imaging subset of DDSM. *Cancer Imaging Arch.* 8, 2016.
- Lee, H., Jo, J., Lim, H., 2020. Study on optimal generative network for synthesizing brain tumor-segmented MR images. *Math. Probl. Eng.* 2020.
- Lee, J., Nishikawa, R.M., 2020. Simulating breast mammogram using conditional generative adversarial network: application towards finding mammographically-occult cancer. In: *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314. International Society for Optics and Photonics, 1131418.
- Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., Aussó, S., Alberich, L.C., Marias, K., Tisknakis, M., et al., 2021. FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv preprint arXiv:2109.09658*.
- Levine, A.B., Peng, J., Farnell, D., Nurse, M., Wang, Y., Naso, J.R., Ren, H., Farahani, H., Chen, C., Chiu, D., et al., 2020. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *J. Pathol.* 252 (2), 178–188.
- Levy, M.A., Rubin, D.L., 2008. Tool support to enable evaluation of the clinical response to treatment. In: *AMIA Annual Symposium Proceedings*, Vol. 2008. American Medical Informatics Association, p. 399.
- Li, X., Cui, Z., Wu, Y., Gu, L., Harada, T., 2021a. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*.
- Li, H., Giger, M.L., Huynh, B.Q., Antropova, N.O., 2017. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *J. Med. Imaging* 4 (4), 041304.
- Li, Y., Han, G., Wu, X., Li, Z.H., Zhao, K., Zhang, Z., Liu, Z., Liang, C., 2021b. Normalization of multicenter CT radiomics by a generative adversarial network method. *Phys. Med. Biol.* 66 (5), 055030.
- Li, T., Lin, L., 2019. Anonymusnet: Natural face de-identification with measurable privacy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al., 2019b. Privacy-preserving federated brain tumour segmentation. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 133–141.
- Li, H., Paetzold, J.C., Sekuboyina, A., Kofler, F., Zhang, J., Kirschke, J.S., Wiestler, B., Menze, B., 2019a. DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 795–803.
- Li, H., Pan, S.J., Wang, S., Kot, A.C., 2018. Domain generalization with adversarial feature learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5400–5409.
- Li, M., Tang, H., Chan, M.D., Zhou, X., Qian, X., 2020a. DC-AL GAN: pseudoprogression and true tumor progression of glioblastoma multimodal image classification based on DCGAN and AlexNet. *Med. Phys.* 47 (3), 1139–1150.
- Li, C., Wand, M., 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In: *European Conference on Computer Vision*. Springer, pp. 702–716.
- Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., Wang, D., 2020b. A large-scale CT and PET/CT dataset for lung cancer diagnosis [dataset]. *Cancer Imaging Arch.*
- Li, Y., Yuan, L., Vasconcelos, N., 2019c. Bidirectional learning for domain adaptation of semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6936–6945.
- Liberman, L., Menell, J.H., 2002. Breast imaging reporting and data system (BI-RADS). *Radiol. Clin.* 40 (3), 409–430.
- Liew, S.-L., Anglin, J.M., Banks, N.W., Sondag, M., Ito, K.L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., et al., 2018. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Sci. Data* 5 (1), 1–11.
- Lim, Z.W., Lee, M.L., Hsu, W., Wong, T.Y., 2019. Building trust in deep learning system towards automated disease detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. pp. 9516–9521.
- Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Hallilovic, A., Hermesen, M., van de Loo, R., Vogels, R., et al., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7 (6), giy065.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88.
- Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med. Image Anal.* 18 (2), 359–373.
- Liu, R., Ge, Y., Choi, C.L., Wang, X., Li, H., 2021. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16377–16386.
- Liu, X., Guo, S., Zhang, H., He, K., Mu, S., Guo, Y., Li, X., 2019a. Accurate colorectal tumor segmentation for CT scans based on the label assignment generative adversarial network. *Med. Phys.* 46 (8), 3532–3542.
- Liu, S., Setio, A.A.A., Ghesu, F.C., Gibson, E., Grbic, S., Georgescu, B., Comaniciu, D., 2020b. No surprises: Training robust lung nodule detection for low-dose ct scans by augmenting with adversarial attacks. *IEEE Trans. Med. Imaging* 40 (1), 335–345.
- Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., Wang, Z., 2019b. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. *Engineering* 5 (1), 156–163.
- Liu, B., Zhu, Y., Song, K., Elgammal, A., 2020a. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In: *International Conference on Learning Representations*.
- Ljosa, V., Sokolnicki, K.L., Carpenter, A.E., 2012a. Annotated high-throughput microscopy image sets for validation. *Nature Methods* 9 (7), 637.
- Ljosa, V., Sokolnicki, K.L., Carpenter, A.E., 2012b. Annotated high-throughput microscopy image sets for validation. *Nature Methods* 9 (7), 637.
- Loeb, S., Bjurlin, M.A., Nicholson, J., Tammela, T.L., Penson, D.F., Carter, H.B., Carroll, P., Etzioni, R., 2014. Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.* 65 (6), 1046–1055.
- Lopez, M., Posada, N., Moura, D.C., Pollán, R.R., Valiente, J.M.F., Ortega, C.S., Solar, M., Diaz-Herrero, G., Ramos, I., Loureiro, J., et al., 2012. BCDR: a breast cancer digital repository. In: *15th International Conference on Experimental Mechanics*, Vol. 1215.
- Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*.
- Lyu, M., Su, D., Li, N., 2016. Understanding the sparse vector technique for differential privacy. *arXiv preprint arXiv:1603.01699*.
- MacKay, D.J., 1992. A practical Bayesian framework for backpropagation networks. *Neural Comput.* 4 (3), 448–472.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Mahajan, V., Venugopal, V., 2020. Audit of artificial intelligence algorithms and its impact in relieving shortage of specialist doctors. In: *Artificial Intelligence: Applications in Healthcare Delivery*. CRC Press, p. 207.
- Mahmood, R., Babier, A., McNiven, A., Diamant, A., Chan, T.C., 2018. Automated treatment planning in radiation therapy using generative adversarial networks. In: *Machine Learning for Healthcare Conference*. PMLR, pp. 484–499.
- Maier, O., Menze, B.H., von der Gabelatz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med. Image Anal.* 35, 250–269.
- Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B., 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

- Mardani, M., Gong, E., Cheng, J.Y., Vasanawala, S., Zaharchuk, G., Alley, M., Thakur, N., Han, S., Dally, W., Pauly, J.M., et al., 2017. Deep generative adversarial networks for compressed sensing automates MRI. *arXiv preprint arXiv:1706.00051*.
- Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kiebert, K., Flagg, E., Chowdhury, S., et al., 2011. The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* 95 (4), 629–635.
- Martin, B., Schäfer, E., Jakubowicz, E., Mayr, P., Irhringer, R., Anthuber, M., Schenkirsch, G., Schaller, T., Märkl, B., 2018. Interobserver variability in the H&E-based assessment of tumor budding in pT3/4 colon cancer: does it affect the prognostic relevance? *Virchows Arch.* 473 (2), 189–197.
- Maspero, M., Savenije, M.H., Dinkla, A.M., Seevinck, P.R., Intven, M.P., Jurgenliemk-Schulz, I.M., Kerkmeijer, L.G., van den Berg, C.A., 2018. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys. Med. Biol.* 63 (18), 185001.
- Mathew, S., Nadeem, S., Kumari, S., Kaufman, A., 2020. Augmenting colonoscopy using extended and directional cyclegan for lossy image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4696–4705.
- Maximov, M., Elezi, I., Leal-Taixé, L., 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456.
- McCreadie, G., Oliver, T., 2009. Eight CT lessons that we learned the hard way: an analysis of current patterns of radiological error and discrepancy with particular emphasis on CT. *Clin. Radiol.* 64 (5), 491–499.
- McDonald, R.J., Schwartz, K.M., Eckel, L.J., Diehn, F.E., Hunt, C.H., Bartholmai, B.J., Erickson, B.J., Kallmes, D.F., 2015. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* 22 (9), 1191–1198.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*. PMLR, pp. 1273–1282.
- Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A., 2021. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*.
- Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34 (10), 1993–2024.
- Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge? In: *International Conference on Machine Learning*. PMLR, pp. 3481–3490.
- Messiou, C., Hillengass, J., Delorme, S., Lecouvet, F.E., Mouloupoulos, L.A., Collins, D.J., Blackledge, M.D., Abildgaard, N., Østergaard, B., Schlemmer, H.-P., et al., 2019. Guidelines for acquisition, interpretation, and reporting of whole-body MRI in myeloma: myeloma response assessment and diagnosis system (MY-RADS). *Radiology* 291 (1), 5–13.
- Milchenko, M., Marcus, D., 2013. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* 11 (1), 65–75.
- Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y., 2019. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. In: *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 461–478.
- Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Modanwal, G., Vellal, A., Mazurowski, M.A., 2019. Normalization of breast MRIs using cycle-consistent generative adversarial networks. *arXiv preprint arXiv:1912.08061*.
- Mok, T.C., Chung, A.C., 2018. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In: *International MICCAI Brainlesion Workshop*. Springer, pp. 70–80.
- Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. *Acad. Radiol.* 19 (2), 236–248.
- Mullick, S.S., Datta, S., Das, S., 2019. Generative adversarial minority oversampling. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1695–1704.
- Munawar, F., Azmat, S., Iqbal, T., Grönlund, C., Ali, H., 2020. Segmentation of lungs in chest X-ray image using generative adversarial networks. *IEEE Access* 8, 153535–153545.
- Murakami, Y., Magome, T., Matsumoto, K., Sato, T., Yoshioka, Y., Oguchi, M., 2020. Fully automated dose prediction using generative adversarial networks in prostate cancer patients. *PLoS One* 15 (5), e0232697.
- Muramatsu, C., Nishio, M., Goto, T., Oiwa, M., Morita, T., Yakami, M., Kubo, T., Togashi, K., Fujita, H., 2020. Improving breast mass classification by shared data with domain transformation using a generative adversarial network. *Comput. Biol. Med.* 119, 103698.
- Nash, J.F., et al., 1950. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci.* 36 (1), 48–49.
- National Cancer Institute, 2018. Radiology data from the clinical proteomic tumor analysis consortium glioblastoma multiforme [cptac-gbm] collection [data set]. *Cancer Imaging Arch.*
- Neal, R.M., 2012. *Bayesian Learning for Neural Networks*, Vol. 118. Springer Science & Business Media.
- Nearchou, I.P., Soutar, D.A., Ueno, H., Harrison, D.J., Arandelovic, O., Caie, P.D., 2021. A comparison of methods for studying the tumor microenvironment's spatial heterogeneity in digital pathology specimens. *J. Pathol. Inform.*
- Negi, A., Raj, A.N.J., Nerissson, R., Zhuang, Z., Murugappan, M., 2020. RDA-UNET-WGAN: An accurate breast ultrasound lesion segmentation using wasserstein generative adversarial networks. *Arab. J. Sci. Eng.* 45 (8), 6399–6410.
- Nehmeh, S., Erdi, Y., Ling, C., Rosenzweig, K., Squire, O., Braban, L., Ford, E., Sidhu, K., Mageras, G., Larson, S., et al., 2002. Effect of respiratory gating on reducing lung motion artifacts in PET imaging of lung cancer. *Med. Phys.* 29 (3), 366–371.
- Network, C.G.A.R., et al., 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* 474 (7353), 609–615.
- Newman-Toker, D.E., Wang, Z., Zhu, Y., Nassery, N., Tehrani, A.S.S., Schaffer, A.C., Yu-Moe, C.W., Clemens, G.D., Fanaei, M., Siegal, D., 2021. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the “Big Three”. *Diagnosis* 8 (1), 67–84.
- Nie, D., Shen, D., 2020. Adversarial confidence learning for medical image segmentation and synthesis. *Int. J. Comput. Vis.* 1–20.
- Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2017. Medical image synthesis with context-aware generative adversarial networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 417–425.
- Nishio, M., Muramatsu, C., Noguchi, S., Nakai, H., Fujimoto, K., Sakamoto, R., Fujita, H., 2020. Attribute-guided image generation of three-dimensional computed tomography images of lung nodules using a generative adversarial network. *Comput. Biol. Med.* 126, 104032.
- NLST Research Team, 2011. The national lung screening trial: overview and study design. *Radiology* 258 (1), 243–253.
- Norton, L., Simon, R., Brereton, H.D., Bogden, A.E., 1976. Predicting the course of Gompertzian growth. *Nature* 264 (5586), 542–545.
- Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier GANs. In: *International Conference on Machine Learning*. PMLR, pp. 2642–2651.
- Oleszkiewicz, W., Kairouz, P., Piczak, K., Rajagopal, R., Trzcinski, T., 2018. Siamese generative adversarial privatizer for biometric data. In: *Asian Conference on Computer Vision*. Springer, pp. 482–497.
- Oliveira, D.A.B., 2020a. Controllable skin lesion synthesis using texture patches, Bézier curves and conditional GANs. In: *2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE*, pp. 1798–1802.
- Oliveira, D.A.B., 2020b. Implanting synthetic lesions for improving liver lesion segmentation in CT exams. *arXiv preprint arXiv:2008.04690*.
- Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K., Fujita, H., 2019. Automated pulmonary nodule classification in computed tomography images using a deep convolutional neural network trained by generative adversarial networks. *BioMed Res. Int.* 2019.
- Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K., Fujita, H., 2020. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. *Int. J. Comput. Assist. Radiol. Surg.* 15 (1), 173–178.
- Pandey, S., Singh, P.R., Tian, J., 2020. An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation. *Biomed. Signal Process. Control* 57, 101782.
- Pang, T., Wong, J.H.D., Ng, W.L., Chan, C.S., 2021. Semi-supervised GAN-based radiomics model for data augmentation in breast ultrasound mass classification. *Comput. Methods Programs Biomed.* 203, 106018.
- Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K., 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.
- Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, U., 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*.
- Park, H., Bayat, A., Sabokrou, M., Kirschke, J.S., Menze, B.H., 2020. Robustification of segmentation models against adversarial perturbations in medical imaging. In: *International Workshop on Predictive Intelligence in Medicine*. Springer, pp. 46–57.
- Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346.
- Park, H., Yoo, Y., Kwak, N., 2018. Mc-gan: Multi-conditional generative adversarial network for image synthesis. *arXiv preprint arXiv:1805.01123*.
- Parmar, C., Grossmann, P., Bussink, J., Lambin, P., Aerts, H.J., 2015. Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* 5 (1), 1–11.
- Patel, P., Kumari, N., Singh, M., Krishnamurthy, B., 2021. Lr-gan: Self-supervised gan with latent transformation detection. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3189–3198.
- Paul, R., Schabath, M., Gillies, R., Hall, L., Goldof, D., 2020. Mitigating adversarial attacks on medical image understanding systems. In: *2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE*, pp. 1517–1521.
- Pawlowski, N., Coelho de Castro, D., Glocker, B., 2020. Deep structural causal models for tractable counterfactual inference. *Adv. Neural Inf. Process. Syst.* 33, 857–869.

- Peng, Y., Chen, S., Qin, A., Chen, M., Gao, X., Liu, Y., Miao, J., Gu, H., Zhao, C., Deng, X., et al., 2020. Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. *Radiother. Oncol.* 150, 217–224.
- Phoulady, H.A., Mouton, P.R., 2018. A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection. *arXiv preprint arXiv:1811.09651*.
- Pinaya, W.H.L., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2021. Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650*.
- Pittaluga, F., Koppal, S., Chakrabarti, A., 2019. Learning privacy preserving encodings through adversarial training. In: 2019 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 791–799.
- Poorneshwaran, J., Kumar, S.S., Ram, K., Joseph, J., Sivaprakasam, M., 2019. Polyp segmentation using generative adversarial network. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 7201–7204.
- Prior, F., Smith, K., Sharma, A., Kirby, J., Tarbox, L., Clark, K., Bennett, W., Nolan, T., Freymann, J., 2017. The public cancer radiology imaging collections of The Cancer Imaging Archive. *Sci. Data* 4 (1), 1–7.
- Prokopenko, D., Stadelmann, J.V., Schulz, H., Renisch, S., Dylov, D.V., 2019. Unpaired synthetic image generation in radiology using GANs. In: Workshop on Artificial Intelligence in Radiation Therapy. Springer, pp. 94–101.
- Pusey, E., Lufkin, R.B., Brown, R., Solomon, M.A., Stark, D.D., Tarr, R., Hanafey, W., 1986. Magnetic resonance imaging artifacts: mechanism and clinical significance. *Radiographics* 6 (5), 891–911.
- Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., Menze, B., 2020. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. In: Medical Imaging with Deep Learning. PMLR, pp. 655–668.
- Qin, Z., Liu, Z., Zhu, P., Xue, Y., 2020. A GAN-based image synthesis method for skin lesion classification. *Comput. Methods Programs Biomed.* 195, 105568.
- Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2009. *Dataset Shift in Machine Learning*. The MIT Press.
- Quiros, A.C., Murray-Smith, R., Yuan, K., 2019. PathologyGAN: Learning deep representations of cancer tissue. *arXiv preprint arXiv:1907.02644*.
- Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S., 2019. Multi-component image translation for deep domain generalization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 579–588.
- Rashid, H., Tanveer, M.A., Khan, H.A., 2019. Skin lesion classification using GAN based data augmentation. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 916–919.
- Rasouli, M., Sun, T., Rajagopal, R., 2020. Fedgan: Federated generative adversarial networks for distributed data. *arXiv preprint arXiv:2006.07228*.
- Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D., 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *Int. J. Comput. Assist. Radiol. Surg.* 14 (7), 1167–1176.
- Raval, N., Machanavajjhala, A., Cox, L.P., 2017. Protecting visual secrets using adversarial nets. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1329–1332.
- Razavi, A., Van den Oord, A., Vinyals, O., 2019. Generating diverse high-fidelity images with vq-vae-2. *Adv. Neural Inf. Process. Syst.* 32.
- Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., Meinel, C., 2017. A conditional adversarial network for semantic segmentation of brain tumor. In: International MICCAI Brainlesion Workshop. Springer, pp. 241–252.
- Rezende, D., Mohamed, S., 2015. Variational inference with normalizing flows. In: International Conference on Machine Learning. PMLR, pp. 1530–1538.
- Rimmer, A., 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: Br. Med. J. (Online)* 359.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
- Rubin, M., Stein, O., Turko, N.A., Nygate, Y., Roitshtain, D., Karako, L., Barnea, I., Giryas, R., Shaked, N.T., 2019. TOP-GAN: Stain-free cancer cell classification using deep learning with a small training set. *Med. Image Anal.* 57, 176–185.
- Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M., 2021a. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*.
- Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M., 2021b. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*.
- Salehinejad, H., Valaei, S., Dowdell, T., Colak, E., Barlett, J., 2018. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 990–994.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. *arXiv preprint arXiv:1606.03498*.
- Samangouei, P., Kabkab, M., Chellappa, R., 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*.
- Samthar, V., Murtua, I., Martín, J.J.A., Gutierrez, A., 2021. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *J. Big Data* 8 (1), 1–59.
- Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M., 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* 9 (1), 1–9.
- Santurkar, S., Schmidt, L., Madry, A., 2018. A classification-based study of covariate shift in GAN distributions. In: International Conference on Machine Learning. PMLR, pp. 4480–4489.
- Sarker, M., Kamal, M., Rashwan, H.A., Abdel-Nasser, M., Singh, V.K., Banu, S.F., Akram, F., Chowdhury, F.U., Choudhury, K.A., Chamon, S., et al., 2019. Mobile-GAN: Skin lesion segmentation using a lightweight generative adversarial network. *arXiv preprint arXiv:1907.00856*.
- Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R., 2018. Fairness GAN. *arXiv preprint arXiv:1805.09910*.
- Sauer, A., Schwarz, K., Geiger, A., 2022. Stylegan-x: Scaling stylegan to large diverse datasets. p. 1, *arXiv preprint arXiv:2202.00273*.
- Schimke, N., Kuehler, M., Hale, J., 2011. Preserving privacy in structural neuroimages. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer, pp. 301–308.
- Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* 54, 30–44.
- Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.
- Schminda, K., Prah, M., 2018. Data from brain-tumor-progression. *Cancer Imaging Arch.*
- Schonfeld, E., Schiele, B., Khoreva, A., 2020. A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8207–8216.
- Schwarz, C.G., Kremers, W.K., Therau, T.M., Sharp, R.R., Gunter, J.L., Vemuri, P., Arani, A., Sychalla, A.J., Kantarci, K., Knopman, D.S., et al., 2019. Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* 381 (17), 1684–1686.
- Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. *Neuroimage* 22 (3), 1060–1075.
- Selvan, R., Faye, F., Middleton, J., Pai, A., 2020. Uncertainty quantification in medical image segmentation with normalizing flows. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 80–90.
- Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med. Image Anal.* 42, 1–13.
- Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., et al., 2014. The Cambridge centre for ageing and neuroscience (cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* 14 (1), 1–25.
- Shaham, T.R., Dekel, T., Michaeli, T., 2019. Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4570–4580.
- Shahidi, F., 2021. Breast cancer histopathology image super-resolution using wide-attention GAN with improved wasserstein gradient penalty and perceptual loss. *IEEE Access* 9, 32795–32809.
- Sharma, N., Aggarwal, L.M., 2010. Automated medical image segmentation techniques. *J. Med. Phys./Assoc. Med. Phys. India* 35 (1), 3.
- Sharpe, M.B., Moore, K.L., Orton, C.G., 2014. Within the next ten years treatment planning will become fully automated without the need for human intervention. *Med. Phys.* 41 (12), 120601. <http://dx.doi.org/10.1118/1.4894496>.
- Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., et al., 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* 10 (1), 1–12.
- Shen, T., Hao, K., Gou, C., Wang, F.-Y., 2021. Mass image synthesis in mammogram with contextual information based on GANs. *Comput. Methods Programs Biomed.* 202, 106019.
- Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248.
- Shi, Z., Hu, Q., Yue, Y., Wang, Z., Al-Othmani, O.M.S., Li, H., 2020. Automatic nodule segmentation method for CT images using aggregation-u-net generative adversarial networks. *Sens. Imaging* 21 (1), 1–16.
- Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* 90 (2), 227–244.

- Shin, Y., Qadir, H.A., Balasingham, I., 2018b. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access* 6, 56007–56017.
- Shin, H.-C., Tenenholz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M., 2018a. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, pp. 1–11.
- Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.* 174 (1), 71–74.
- Shokri, R., Shmatikov, V., 2015. Privacy-preserving deep learning. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. pp. 1310–1321.
- Shyamala, K., Girish, H., Murgod, S., 2014. Risk of tumor cell seeding through biopsy and aspiration cytology. *J. Int. Soc. Prev. Community Dent.* 4 (1), 5.
- Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J., 2019. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 191–200.
- Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* 9 (2), 283–293.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.
- Singh, S., Bray, M.A., Jones, T.R., Carpenter, A.E., 2014. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.* 256 (3), 231–236.
- Singh, N.K., Raza, K., 2020. Medical image generation using generative adversarial networks. *arXiv preprint arXiv:2005.10687*.
- Singh, V.K., Romani, S., Rashwan, H.A., Akram, F., Pandey, N., Sarker, M.M.K., Abdulwahab, S., Torrents-Barrena, J., Saleh, A., Arquez, M., et al., 2018. Conditional generative adversarial and convolutional networks for X-ray breast mass segmentation and shape classification. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 833–840.
- Sogancioglu, E., Murphy, K., van Ginneken, B., 2021. NODE21. <http://dx.doi.org/10.5281/zenodo.5548363>.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265.
- Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J., 3D Image Reconstruction for Comparison of Algorithm Database. URL: <https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircad-b01/>.
- Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* 32.
- Sorin, V., Barash, Y., Konen, E., Kiang, E., 2020. Creating artificial images for radiology applications using generative adversarial networks (GANs)—a systematic review. *Acad. Radiol.*
- Stadler, T., Oprisanu, B., Troncoso, C., 2021. Synthetic data – anonymisation groundhog day. *arXiv:2011.07018*.
- Stein, C.M., 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 1135–1151.
- Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, pp. 240–248.
- Sun, B., Liu, F., Zhou, Y., Jin, S., Li, Q., Jin, X., 2020a. Classification of lung nodules based on GAN and 3D CNN. In: *Proceedings of the 4th International Conference on Computer Science and Application Engineering*. pp. 1–5.
- Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J., 2020b. An adversarial learning approach to medical image synthesis for lesion detection. *IEEE J. Biomed. Health Inf.* 24 (8), 2303–2314.
- Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A., 2020. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*.
- Swiderski, B., Gielata, L., Olszewski, P., Osowski, S., Kolodziej, M., 2021. Deep neural system for supporting tumor recognition of mammograms using modified GAN. *Expert Syst. Appl.* 164, 113968.
- Swieciecki, A., Konz, N., Buda, M., Mazurowski, M.A., 2021. A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis. *Sci. Rep.* 11 (1), 1–13.
- Szafranowska, Z., Osuala, R., Breier, B., Kushibar, K., Lekadir, K., Diaz, O., 2022. Sharing generative models instead of private data: a simulation study on mammography patch classification 12286. pp. 169–177. <http://dx.doi.org/10.1117/12.2625781>.
- Tang, Y.-B., Tang, Y.-X., Xiao, J., Summers, R.M., 2019. Xlsor: A robust and accurate lung segmentor on chest X-Rays using criss-cross attention and customized radio-realistic abnormalities generation. In: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 457–467.
- Tanner, C., Ozdemir, F., Profanter, R., Vishnevsky, V., Konukoglu, E., Goksel, O., 2018. Generative adversarial networks for mr-ct deformable image registration. *arXiv preprint arXiv:1807.07349*.
- Tanno, R., Worrall, D.E., Kaden, E., Ghosh, A., Grussu, F., Bizzi, A., Sotiropoulos, S.N., Criminisi, A., Alexander, D.C., 2021. Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI. *NeuroImage* 225, 117366.
- Teh, S., Rangui, S., Fagan, P., 2017. Inter-observer variability between radiologists reporting on cerebellopontine angle tumours on magnetic resonance imaging. *J. Laryngol. Otol.* 131 (S1), S47–S49.
- Teramoto, A., Tsukamoto, T., Yamada, A., Kiriya, Y., Imaizumi, K., Saito, K., Fujita, H., 2020. Deep learning approach to classification of lung cytological images: Two-step training using actual and synthesized images by progressive growing of generative adversarial networks. *PLoS One* 15 (3), e0229951.
- Thambawita, V., Salehi, P., Sheshkal, S.A., Hicks, S.A., Hammer, H.L., Parasa, S., Lange, T.D., Halvorsen, P., Riegler, M.A., 2022. SinGAN-Seg: Synthetic training data generation for medical image segmentation. *PLoS One* 17 (5), e0267976.
- Tomczak, K., Czerwińska, P., Wizerowicz, M., 2015. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* 19 (1A), 68–77.
- Torfi, A., Fox, E.A., Reddy, C.K., 2020. Differentially private synthetic medical data generation using convolutional GANs. *arXiv preprint arXiv:2012.11774*.
- Jimenez-del Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A.A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., et al., 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Trans. Med. Imaging* 35 (11), 2459–2475.
- Trebb, B.E., Cox, B.T., 2010. k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields. *J. Biomed. Opt.* 15 (2), 021314.
- Troyanskaya, O., Trajanoski, Z., Carpenter, A., Thrun, S., Razavian, N., Oliver, N., 2020. Artificial intelligence and cancer. *Nature Cancer* 1 (2), 149–152.
- Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* 5 (1), 1–9.
- Tschuchnig, M.E., Oostingh, G.J., Gadermayr, M., 2020. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns* 1 (6), 100089.
- Vallières, M., Kayrivest, E., Perrin, L., et al., 2017. Data from head-neck-pet-ct. *Cancer Imaging Arch.*
- Van Den Oord, A., Vinyals, O., et al., 2017. Neural discrete representation learning. *Adv. Neural Inf. Process. Syst.* 30.
- Van der Goten, L.A., Hepp, T., Akata, Z., Smith, K., 2021. Adversarial privacy preservation in MRI scans of the brain. URL: <https://openreview.net/forum?id=2NHI-ETnHxk>.
- Van Tulder, G., de Bruijne, M., 2015. Why does synthesized data improve multi-sequence classification? In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 531–538.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdal, M., Courville, A., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* 2017.
- Verma, N., Cowperthwaite, M.C., Burnett, M.G., Markey, M.K., 2013. Differentiating tumor recurrence from treatment necrosis: a review of neuro-oncologic imaging strategies. *Neuro-Oncol.* 15 (5), 515–534.
- Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. pp. 1096–1103.
- Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., Savarese, S., 2018. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*.
- Vu, Q.D., Kim, K., Kwak, J.T., 2020a. Unsupervised tumor characterization via conditional generative adversarial networks. *IEEE J. Biomed. Health Inf.*
- Vu, T., Li, M., Humayun, H., Zhou, Y., Yao, J., 2020b. A generative adversarial network for artifact removal in photoacoustic computed tomography with a linear-array transducer. *Exp. Biol. Med.* 245 (7), 597–605.
- Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13 (4), 600–612.
- Wang, T., Lei, Y., Curran, W.J., Liu, T., Yang, X., 2021. Contrast-enhanced MRI synthesis from non-contrast MRI using attention CycleGAN. In: *Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging*, Vol. 11600. International Society for Optics and Photonics, 116001L.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018a. High-resolution image synthesis and semantic manipulation with conditional GANs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8798–8807.

- Wang, Z., She, Q., Ward, T.E., 2019b. Generative adversarial networks in computer vision: A survey and taxonomy. *arXiv preprint arXiv:1906.01529*.
- Wang, H., Shen, T., Zhang, W., Duan, L.-Y., Mei, T., 2020a. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: *European Conference on Computer Vision*. Springer, pp. 642–659.
- Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., Shoemaker, B.A., Gindulyte, A., Bryant, S.H., 2014. *PubChem BioAssay*: 2014 update. *Nucl. Acids Res.* 42 (Database issue), D1075–1082.
- Wang, Q., Zhang, X., Chen, W., Wang, K., Zhang, X., 2020b. Class-aware multi-window adversarial lung nodule synthesis conditioned on semantic features. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 589–598.
- Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., Ordonez, V., 2019a. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5310–5319.
- Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M., 2022. Diffusion-GAN: Training GANs with diffusion. <http://dx.doi.org/10.48550/ARXIV.2206.02262>, URL: <https://arxiv.org/abs/2206.02262>.
- Wang, Y., Zhou, L., Wang, M., Shao, C., Shi, L., Yang, S., Zhang, Z., Feng, M., Shan, F., Liu, L., 2020c. Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification. *Quant. Imaging Med. Surg.* 10 (6), 1249.
- Wang, Y., Zhou, L., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., Shen, D., 2018b. 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. *IEEE Trans. Med. Imaging* 38 (6), 1328–1339.
- Wei, L., Lin, Y., Hsu, W., 2020. Using a generative adversarial network for CT normalization and its impact on radiomic features. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 844–848.
- Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, Jr., C.R., Jagust, W., Morris, J.C., et al., 2017. The Alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's Dement.* 13 (5), 561–571.
- Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genet.* 45 (10), 1113–1120.
- Wetstein, S.C., González-Gonzalo, C., Bortsova, G., Liefers, B., Dubost, F., Katramados, I., Hogeweg, L., van Ginneken, B., Pluim, J.P., de Bruijne, M., et al., 2020. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *arXiv preprint arXiv:2006.06356*.
- Wilson, M.K., Chawla, T., Wang, L., Friedlander, M., Oza, A.M., Lheureux, S., 2018. Inter and intra-observer variability with the assessment of RECIST in ovarian cancer.
- Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, pp. 14–23.
- Wolterink, J.M., Kamnitsas, K., Ledig, C., Išgum, I., 2018. Generative adversarial networks and adversarial methods in biomedical image analysis. *arXiv preprint arXiv:1810.10352*.
- Woo, M., Lowe, S.C., Devane, A.M., Gimbel, R.W., 2020. Intervention to reduce interobserver variability in computed tomographic measurement of cancer lesions among experienced radiologists. *Curr. Probl. Diagn. Radiol.*
- World Health Organization, 2018. Cancer. URL: <https://www.who.int/news-room/factsheets/detail/cancer>.
- Wu, Z., Wang, Z., Wang, Z., Jin, H., 2018c. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 606–624.
- Wu, E., Wu, K., Cox, D., Lotter, W., 2018a. Conditional infilling GANs for data augmentation in mammogram classification. In: *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, pp. 98–106.
- Wu, E., Wu, K., Lotter, W., 2020. Synthesizing lesions using contextual GANs improves breast cancer classification on mammograms. *arXiv preprint arXiv:2006.00086*.
- Wu, Y., Yang, F., Ling, H., 2018b. Privacy-protective-GAN for face de-identification. *arXiv preprint arXiv:1806.08906*.
- Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al., 2013. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's Dement.* 9 (3), 332–337.
- Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., Alzheimer's Disease Neuroimaging Initiative, et al., 2021. Learning to synthesise the ageing brain without longitudinal data. *Med. Image Anal.* 73, 102169.
- Xiao, X., Zhao, J., Qiang, Y., Chong, J., Yang, X., Kazihise, N.G.-F., Chen, B., Li, S., 2019. Radiomics-guided GAN for segmentation of liver tumor without contrast agents. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 237–245.
- Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y., 2020. MI2GAN: Generative adversarial network for medical image domain adaptation using mutual information constraint. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 516–525.
- Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J., 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.
- Xin, B., Yang, W., Geng, Y., Chen, S., Wang, S., Huang, L., 2020. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2927–2931.
- Xu, Z., Wang, X., Shin, H.-C., Yang, D., Roth, H., Milletari, F., Zhang, L., Xu, D., 2020. Correlation via synthesis: End-to-end image generation and radiogenomic learning based on generative adversarial network. In: *Medical Imaging with Deep Learning*. PMLR, pp. 857–866.
- Xu, D., Yuan, S., Zhang, L., Wu, X., 2018. Fairgan: Fairness-aware generative adversarial networks. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 570–575.
- Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SEGAN: Adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics* 16 (3), 383–392.
- Yan, K., Wang, X., Lu, L., Summers, R.M., 2018. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J. Med. Imaging* 5 (3), 036501.
- Yang, T.-Y., Brinton, C., Mittal, P., Chiang, M., Lan, A., 2018c. Learning informative and private representations via generative adversarial networks. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1534–1543.
- Yang, J., Liu, S., Grbic, S., Setio, A.A.A., Xu, Z., Gibson, E., Chabin, G., Georgescu, B., Laine, A.F., Comaniciu, D., 2019. Class-aware adversarial lung nodule synthesis in CT images. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, pp. 1348–1352.
- Yang, J., Veeraraghavan, H., Armato, III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmp, W., Dekker, A., Han, X., Feng, X., et al., 2018b. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. *Med. Phys.* 45 (10), 4568–4581.
- Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., Firmin, D., 2018a. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging* 37 (6), 1310–1321. <http://dx.doi.org/10.1109/TMI.2017.2785879>.
- Yi, X., Wallia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* 58, 101552.
- Yoon, J., Drumright, L.N., Van Der Schaar, M., 2020. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE J. Biomed. Health Inf.* 24 (8), 2378–2388.
- Yoon, J., Jordan, J., Schaar, M., 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: *ICLR*.
- You, C., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z., Cong, W., et al., 2019. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). *IEEE Trans. Med. Imaging* 39 (1), 188–203.
- Yu, X., Cai, X., Ying, Z., Li, T., Li, G., 2018b. SingleGAN: Image-to-image translation by a single-generator network using multiple generative adversarial learning. In: *Asian Conference on Computer Vision*. Springer, pp. 341–356.
- Yu, H., Hong, S., Yang, X., Ni, J., Dan, Y., Qin, B., 2013. Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. *BioMed Res. Int.* 2013.
- Yu, H., Zhang, X., 2020. Synthesis of prostate MR images for classification using capsule network-based GAN model. *Sensors* 20 (20), 5736.
- Yu, B., Zhou, L., Wang, L., Frapp, J., Bourgeat, P., 2018a. 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 626–630.
- Yu, B., Zhou, L., Wang, L., Shi, Y., Frapp, J., Bourgeat, P., 2019. Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. *IEEE Trans. Med. Imaging* 38 (7), 1750–1762.
- Yurt, M., Dar, S.U.H., Erdem, A., Erdem, E., Çukur, T., 2019. muSTGAN: Multi-stream generative adversarial networks for MR image synthesis. *arXiv preprint arXiv:1909.11504*.
- Zaman, A., Park, S.H., Bang, H., Park, C.-w., Park, I., Joong, S., 2020. Generative approach for data augmentation for deep learning-based bone surface segmentation from ultrasound images. *Int. J. Comput. Assist. Radiol. Surg.* 15, 931–941.
- Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019a. Self-attention generative adversarial networks. In: *International Conference on Machine Learning*. PMLR, pp. 7354–7363.
- Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018c. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 586–595.
- Zhang, B.H., Lemoine, B., Mitchell, M., 2018a. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 335–340.
- Zhang, Y., Qu, H., Chang, Q., Liu, H., Metaxas, D., Chen, C., 2021. Training federated GANs with theoretical guarantees: A universal aggregation approach. *arXiv preprint arXiv:2102.04655*.
- Zhang, Q., Wang, H., Lu, H., Won, D., Yoon, S.W., 2018b. Medical image synthesis with generative adversarial networks for tissue recognition. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, pp. 199–207.

- Zhang, H., Zhang, Z., Odena, A., Lee, H., 2019b. Consistency regularization for generative adversarial networks. arXiv preprint [arXiv:1910.12027](https://arxiv.org/abs/1910.12027).
- Zhang, Z., Zhao, T., Gay, H., Sun, B., Zhang, W., 2020. Semi-supervised semantic segmentation of organs at risk on 3D pelvic CT images. arXiv preprint [arXiv:2009.09571](https://arxiv.org/abs/2009.09571).
- Zhao, J., Li, D., Kassam, Z., Howey, J., Chong, J., Chen, B., Li, S., 2020a. TripartiteGAN: synthesizing liver contrast-enhanced MRI to improve tumor detection. *Med. Image Anal.* 63, 101667.
- Zhao, B., Tan, Y., Tsai, W.-Y., Qi, J., Xie, C., Lu, L., Schwartz, L.H., 2016. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* 6 (1), 1–7.
- Zhao, K., Zhou, L., Gao, S., Wang, X., Wang, Y., Zhao, X., Wang, H., Liu, K., Zhu, Y., Ye, H., 2020b. Study of low-dose PET image recovery using supervised learning with CycleGAN. *PLoS One* 15 (9), e0238455.
- Zhao, D., Zhu, D., Lu, J., Luo, Y., Zhang, G., 2018. Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16. *Symmetry* 10 (10), 519.
- Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L., 2020. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. *IEEE Trans. Med. Imaging* 39 (9), 2772–2781.
- Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M., 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* 1–19.
- Zhou, M., Leung, A., Echegaray, S., Gentles, A., Shrager, J.B., Jensen, K.C., Berry, G.J., Plevritis, S.K., Rubin, D.L., Napel, S., et al., 2018. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology* 286 (1), 307–315.
- Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C., 2021. Domain generalization: A survey. arXiv preprint [arXiv:2103.02503](https://arxiv.org/abs/2103.02503).
- Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2223–2232.

2.1. Contribution: Medical Image Analysis 2023 (Review & Meta-Analysis)

Chapter 3

Synthesis of Multi-Sequence Dynamic Contrast-Enhanced Breast MRI using Generative Adversarial Networks

This chapter investigates the simulation of breast MRI contrast enhancement without the need for physical contrast agents. Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) is a crucial tool for breast cancer diagnosis and treatment planning. However, the use of contrast agents, such as gadolinium, poses risks such as allergic reactions, kidney issues, and environmental contamination. The studies in this chapter introduce, design, implement, and evaluate conditional generative adversarial networks to generate, based on a pre-contrast input image, corresponding synthetic images for either a single or multiple temporal DCE-MRI sequences. We further define the dimensions of synthetic data evaluation and translate them into concrete validation measures to assess our generated DCE-MRI in a principled manner. Simulating contrast manifestations on DCE-MRI not only enables unsupervised detection of cancerous lesions, but also has the potential to reduce the dependency on physical contrast agents, making the process safer and faster, especially for patient groups who cannot undergo contrast-enhanced imaging.

3.1 Contribution: Simulating Dynamic Tumor Contrast Enhancement in Breast MRI using Conditional Generative Adversarial Networks

Summary:

This paper explores GANs to predict and simulate breast MRI contrast enhancement. This work proposes a framework using conditional GANs that can transform non-contrast T1-weighted fat-saturated MRI images into synthetic DCE-MRI images. Extending on previous work of the dissertation author [166], this model is further trained to generate multiple DCE-MRI timepoints, simulating the dynamics of contrast enhancement over time. The study assesses model performance and contrast enhancement patterns through a variety of experiments, showing that GAN-generated synthetic images are effective in localizing, detecting, and segmenting breast cancer tumors. Additionally, the study introduces the Scaled Aggregate Measure (SAME) to assess and select the best synthetic data generators based on a combination of several complementary metrics.

The results demonstrate that synthetic data generated by GANs can improve breast tumor segmentation models and make these models more robust, particularly when pre-contrast images are used in scenarios where contrast agent administration is not possible. This approach holds promise for reducing the reliance on physical contrast agents in clinical practice while maintaining accurate diagnostic capabilities.

Main Contributions:

- **Virtual Contrast Injection Framework:** Introduces a GAN-based framework capable of simulating DCE-MRI contrast enhancement from pre-contrast MRI data without thereby providing a non-invasive, cost-effective alternative to gadolinium-based contrast agents.
- **Multiple Timepoint DCE-MRI Generation:** Extends the GAN model to simulate contrast enhancement across multiple temporal DCE-MRI sequences, providing insight into dynamic tumor behavior.
- **Scaled Aggregate Measure (SAME):** Proposes the SAME, a unified metric for evaluating the quality of synthetic images by combining several metrics such as structural similarity and perceptual differences.

Chapter 3. Synthesis of Multi-Sequence Dynamic Contrast-Enhanced Breast MRI using Generative Adversarial Networks

- **Downstream Task Validation:** Demonstrates that GAN-generated synthetic DCE-MRI data enhances the robustness of breast tumor segmentation models in situations where domain shifts occur (e.g., between pre- and post-contrast data).
- **Synthetic Data Evaluation Framework:** The key components of desirable characteristics of synthetic data are defined to further qualitatively and quantitatively validate the synthetic DCE-MRI sequences including an analysis of their temporal contrast enhancement patterns.

Contributions | The author of the dissertation led the conceptualization and development of the study, including the ideation, design, implementation, validation and preparation of the manuscript. Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W.H.L., Lang, D.M., Schnabel, J.A., Diaz, O., Lekadir, K. provided critical feedback and assisted in manuscript preparation. Joshi, S. and Garrucho, L. helped in the curation of the annotations of the dataset. Joshi, S. assisted in the implementation of the segmentation experiments. Tsirikoglou, A. assisted in the visualization of the multi-sequence DCE-MRI synthesis results.

Declaration of Doctoral Use | None of the co-authors have used, either implicitly or explicitly, this publication for the development of their own doctoral thesis, nor do they intend to do so in the future.

Source Code | The reproducible source code of this work is made readily available at <https://github.com/RichardObi/SimulatingDCE>.

Copyright Statement | This work is currently submitted and under review as: "Osuala, R., Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W.H.L., Lang, D.M., Schnabel, J.A., Diaz, O., Lekadir, K. (2024). Simulating Dynamic Tumor Contrast Enhancement in Breast MRI using Conditional Generative Adversarial Networks. Submitted to: *Journal of Medical Imaging* (Under Review)."

<https://doi.org/10.48550/arXiv.2409.18872> (Preprint)

Simulating Dynamic Tumor Contrast Enhancement in Breast MRI using Conditional Generative Adversarial Networks

Richard Osuala^{a, b, c, *}, Smriti Joshi^a, Apostolia Tsirikoglou^d, Lidia Garrucho^{a, d}, Walter H. L. Pinaya^e, Daniel M. Lang^{b, c}, Julia A. Schnabel^{b, c, e}, Oliver Diaz^{a, f, **}, Karim Lekadir^{a, g, **}

^aBarcelona Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

^bInstitute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Munich, Germany

^cSchool of Computation, Information and Technology, Technical University of Munich, Munich, Germany

^dDepartment of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden

^eSchool of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom

^fComputer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain

^gInstitució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain

Abstract.

Purpose: Deep generative models and synthetic data generation have become essential for advancing computer-assisted diagnosis and treatment. We explore one such emerging and particularly promising application of deep generative models, namely, the generation of virtual contrast enhancement. This allows to predict and simulate contrast enhancement in breast magnetic resonance imaging (MRI) without physical contrast agent injection, thereby unlocking lesion localization and categorization even in patient populations where the lengthy, costly and invasive process of physical contrast agent injection is contraindicated.

Approach: We define a framework for desirable properties of synthetic data, which leads us to propose the Scaled Aggregate Measure (SAME) consisting of a balanced set of scaled complementary metrics for generative model training and convergence evaluation. We further adopt a conditional generative adversarial network to translate from non-contrast-enhanced T1-weighted fat-saturated breast MRI slices to their dynamic contrast-enhanced (DCE) counterparts, thus learning to detect, localize, and adequately highlight breast cancer lesions. Next, we extend our model approach to jointly generate multiple DCE-MRI timepoints, enabling the simulation of contrast enhancement across temporal DCE-MRI acquisitions. Additionally, 3D U-Net tumor segmentation models are implemented and trained on combinations of synthetic and real DCE-MRI data to investigate the effect of data augmentation with synthetic DCE-MRI volumes.

Results: Conducting four main sets of experiments, (i) the variation across single metrics demonstrated the value of SAME, and (ii) the quality and potential of virtual contrast injection for tumor detection and localization were shown. Segmentation models (iii) augmented with synthetic DCE-MRI data were more robust in the presence of domain shifts between pre-contrast and DCE-MRI domains. The joint synthesis approach of multi-sequence DCE-MRI (iv) resulted in temporally coherent synthetic DCE-MRI sequences and indicated the generative model's capability of learning complex contrast enhancement patterns.

Conclusion: Virtual contrast injection can result in accurate synthetic DCE-MRI images, potentially enhancing breast cancer diagnosis and treatment protocols. We demonstrate that detecting, localizing, and segmenting tumors using synthetic DCE-MRI is feasible and promising, particularly considering patients where contrast agent injection is risky or contraindicated. Jointly generating multiple subsequent DCE-MRI sequences can increase image quality and unlock clinical applications assessing tumor characteristics related to its response to contrast media injection as a pillar for personalized treatment planning.

Keywords: Contrast Agent, Breast Cancer, DCE-MRI, Generative Models, Synthetic Data.

*Corresponding author: Richard Osuala, Richard.Osuala@ub.edu

**These authors contributed equally to this work.

1 Introduction

Deep learning progress in breast cancer imaging Breast cancer was the most common cancer diagnosis worldwide in 2020, taking people of all ages and genders into account. The staggering number of 2.26 million new cases and 684,996 reported deaths underline the significant global burden of breast cancer.¹ In breast cancer imaging and medical imaging at large, deep learning has been gaining popularity due to its promising capabilities of sifting through image data to uncover hidden associations. This capacity allows trained deep learning models to recognize subtle patterns in unseen data, which enables solving a plethora of clinical tasks with high potential to improve patient care. With the emergence of deep learning, vast progress has been observed, e.g., in the promising development of automatic methods for the screening, diagnosis, treatment, and monitoring of cancer based on dynamic contrast enhanced magnetic resonance imaging (DCE-MRI). Such methods include the automated tumor detection, localization, segmentation, and characterization for preoperative planning, patient survival assessment, quantification of recurrence risk, and estimation of treatment response.²⁻⁷

Usage and benefit of contrast agents Through the alteration of magnetic properties of tissue, intravenously administered contrast agents (CA), which are commonly based on gadolinium (Gd), manifest as hyper-intense in DCE-MRI. Thus, they allow the visualization of blood flow and changes in permeability. Multiple DCE-MRI volumes are consecutively acquired (before, during, and after CA administration) to enable the time-dependent evaluation of tissue characteristics and assessment of potential abnormalities. The time-signal intensity curves of such dynamic contrast sequences reflect signal intensity changes induced by the uptake and wash-out of CA over time.⁸⁻¹⁰ DCE-MRI contrast uptake serves as an important biomarker in oncology, enabling cancer detection, characterization, subtype determination, differentiation of malignancy, recurrence prediction, and treatment response assessment.^{2,11,12} Notably, kinetic analysis of contrast enhancement in breast DCE-MRI plays a crucial role in lesion characterization, with features such as peak enhancement, time-to-peak, wash-in and wash-out slopes reflecting malignancy risk. Furthermore, DCE-MRI reveals temporal patterns of contrast enhancement that are not only correlated with breast cancer presence but also offer insights into genetic alterations associated with risk of recurrence, response to chemotherapy, and even the underlying molecular subtypes of breast tumors.^{8-10,12,13}

Disadvantages of contrast agents Despite their undeniable value in diagnostic imaging, gadolinium-based contrast agents (GBCAs) involve concerns regarding their safety profile. For instance, GBCA administration has been linked to an increased risk of nephrogenic systemic fibrosis (NSF).¹⁴ A further concern is given by the deposition of residual gadolinium and its potential bioaccumulation within the body with unknown long-term clinical significance.¹⁴⁻¹⁸ Following a 2016 European Commission request for a review of GBCAs, the European Medicines Agency (EMA) recommended restrictions on specific intravenous linear agents to mitigate potential risks associated with gadolinium deposition.¹⁹ Apart from that, safety concerns extend beyond deposition and NSF, as they also comprise well-known acute effects, such as physiologic and allergic-like reactions, as well as Symptoms Associated with Gadolinium Exposure (SAGE) for which the causal relationship to GBCA is still unknown.²⁰ Furthermore, the administration process involves various drawbacks, including lengthy protocols and scan times, significant financial costs, and the requirement for intravenous cannulation and injection procedures. The reliance of DCE-MRI on

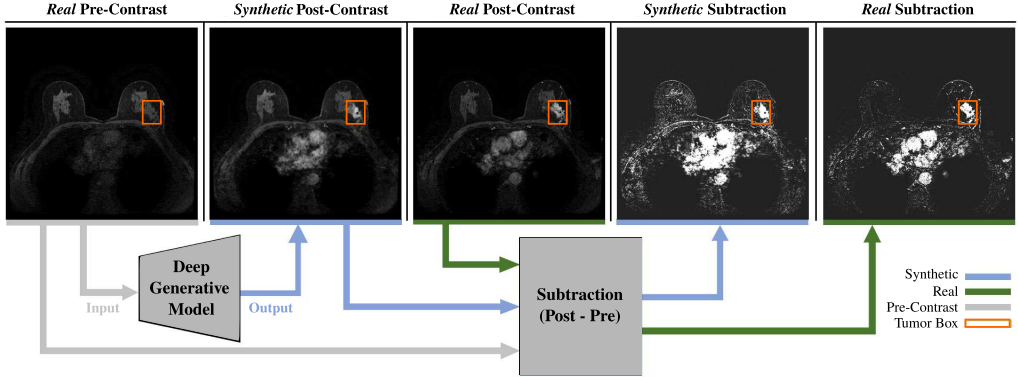


Fig 1: Overview of pre- to post-contrast DCE-MRI synthesis using deep generative models, thereby localizing the contrast-enhanced tumor. Extending single sequence to multi-sequence DCE-MRI image generation further allows the characterization of tumors based on their temporal patterns of contrast agent uptake. The resulting synthetic images can be added as training data for downstream tasks (e.g., tumor segmentation), but, as shown, they can also be utilized to compute subtraction images commonly used in clinical settings for the diagnosis and treatment of breast cancer.

multiple temporal acquisitions further exacerbates the increased costs and extended examination times for patients. Additionally, susceptibility to motion artifacts necessitates meticulous patient cooperation (e.g., breath-holding) and can be a source of discomfort. Collectively, the aforementioned issues contribute to an undue burden on patients, encompassing both inconvenience and potential risks to well-being.^{15,18} Additionally, GBCA administration has been causing the emergence of Gadolinium as a contaminant polluting aquatic ecosystems and environments^{21,22} including drinking water supplies, where its respective degradation products can further increase the risk of adverse health effects.²³ As a consequence, GBCA administration is contraindicated in multiple scenarios, which include patient populations with adverse reactions, pregnancy, kidney malfunctions, missing consent, or high-risk breast cancer screening populations where GBCA exposure extends recommended thresholds in accumulated dosage or frequency.^{24–26}

Contribution The aforementioned issues emphasize the necessity of alternatives that can be safely issued while also simulating GBCA administration in a way that still yields some of its benefits. To this end, we propose the synthetic generation of DCE-MRI using deep generative models, which constitutes a faster, motion artifact-free, and non-invasive alternative with improved cost-effectiveness that also avoids burdening patient health and well-being. Extending over *Osuala et al. (2024)*,²⁵ we provide a principled definition of trustworthy synthetic data and introduce the simultaneous generation of images from multiple DCE-MRI timepoints. We further include a quantitative and qualitative evaluation for jointly generated several DCE-MRI timepoints. Additionally, the temporal contrast enhancement is analyzed at the lesion level including a quantification of contrast intensity patterns per patient case and accumulated over the entire dataset. Overall, our work presents the following contributions and novelties to advance the field of synthetic DCE-MRI for

breast cancer applications:

- *Pre-contrast to DCE-MRI synthesis:* We implement and validate a conditional generative adversarial image synthesis model capable of translating pre-contrast to DCE breast MRI axial slices, which includes lesion detection, localization, and realistic contrast manifestation.
- *Generative model selection framework:* We provide a principled definition of trustworthy synthetic data upon which we derive the Scaled Aggregate Measure (SAME) and validate it by finding the optimal synthetic data generator. SAME combines perceptual and pixel-level synthetic data evaluation, enabling comparisons across generative models and the selection of optimal training checkpoints.
- *Clinical utility validation of synthetic DCE-MRI volumes for tumor segmentation:* We demonstrate the potential of our synthetic data by incorporating it into breast tumor segmentation pipelines. This approach enhances robustness across data domains by providing a wider range of training data while showing the coherence of our synthetic axial slices when stacked as 3D MRI volumes.
- *Joint synthesis of multiple DCE-MRI timepoints:* We introduce and empirically validate the joint generation of images from multiple DCE-MRI timepoints using GANs. Further, intensity distributions are extracted from the region-of-interest to model and assess contrast enhancement patterns of real and synthetic data on individual and dataset level.

2 Related Work

Generative models such as generative adversarial networks (GANs)²⁷ and denoising probabilistic diffusion models (DDPMs)^{28–30} and Latent Diffusion Models (LDMs)³¹ have been widely applied to medical imaging in general and breast imaging in particular.^{4,32–34} For example, *Khader et al. (2023)*³⁵ use unconditional DDPMs to generate non-fat saturated T1-weighted breast DCE-MRI sequences. A set of models have been proposed to condition the generation process on input images^{36–42} unlocking image-to-image translation and domain-adaptation applications in medical imaging.^{43–45} *Konz et al. (2024)*,⁴⁶ for instance, conditioned LDMs on anatomical segmentation masks to generate pre-contrast breast MRI based on 100 patient cases from the Duke-Breast-Cancer-MRI Dataset.⁴⁷

Furthermore, a few first studies started to condition generative models on pre-contrast images to generate their post-contrast counterparts.^{24,26,48–50} For instance, *Kim et al. (2022)*⁵¹ designed a tumor-attentive segmentation-guided GAN that synthesizes a contrast-enhanced T1 breast MRI image from a pre-contrast image, while being guided by the predictions of a surrogate segmentation network. However, with the objective of improving segmentation using GAN-generated data, it can become counterproductive to limit the GAN contrast translation to the tumor segmentation predicted by the segmentation model. Similarly, but based on a chained tumor detection model instead of a segmentation model, *Zhao et al. (2020)*⁵² introduced Tripartite-GAN to generate contrast-enhanced from non contrast-enhanced liver MRI. As high-quality annotations such as segmentation masks or region-of-interest bounding boxes are costly to annotate and, therefore, likely a scarce resource,⁴ it is desirable to accomplish the task of pre- to post-contrast synthesis without relying on such annotations. *Wang et al. (2021)*⁵³ introduced a two-stage GAN that, in its first stage segments the contrast enhancement of the T1-weighted image based on an adversarial loss. Next, in its second stage, it is trained to generate post-contrast DCE images relying on

the segmentation network from the first stage using an L1-loss, an adversarial loss, and an edge detector based L2 loss. *Xue et al. (2022)*⁴⁹ presented a pre- to post-contrast and post- to pre-contrast GAN for brain MRI images. Their bi-directional GAN encodes contrast and image in separate latent representations with the contrast representation producing a contrast enhancement map as output, which can then be subtracted from the synthetic post-contrast image to create the corresponding pre-contrast image.

Recent work has further demonstrated the potential of virtual contrast agents across various medical imaging modalities.^{54–61} For instance, a feasibility study explored replacing gadolinium with a Bayesian deep learning model that predicts virtual contrast enhancement from non-contrast multiparametric brain MRI.⁶² Similarly, a multicenter, retrospective neuro-oncology study showed that convolutional neural networks can generate synthetic post-contrast T1-weighted MRI from pre-contrast scans, enabling accurate tumor burden assessment comparable to real contrast-enhanced images.⁶³

In parallel, novel methodological advances have emerged. One example is the Conditional Autoregressive Vision Model (CAVM), which synthesizes contrast-enhanced brain MRI using masked self-attention in an autoregressive framework that simulates progressive contrast agent dosage.⁶⁴ Another approach leverages conditional diffusion and flow-matching models to incorporate uncertainty estimation in virtual contrast synthesis.⁶⁵

Further, diffusion models have been shown to harmonize varying contrast levels while maintaining consistent segmentation performance in brain MRI.⁶⁶ Conditional GANs have also been trained on scans acquired at varying GBCA doses to synthesize contrast-enhanced images at fractional levels, using a patch-based Wasserstein loss to preserve noise characteristics.^{67,68} Other works have introduced novel loss functions to improve synthesis fidelity, such as frequency-domain consistency⁶⁹ and tumor-focused learning objectives.⁵⁰

*Müller-Franzes et al. (2023)*⁴⁸ translated T1 and T2 images to contrast-enhanced breast MRI images using pix2pixHD⁷⁰ and conducted an observer study to test the realism of the synthetic images. Furthermore, *Müller-Franzes et al. (2024)*⁷¹ compare in a more recent work Diffusion Models and GANs for synthesizing higher-dose DCE-breast MRI subtraction images from their lower-dose counterparts. *Han et al. (2023)*⁷² model the translation of Diffusion Weighted Imaging (DWI) from breast DCE-MRI volumes as sequence-to-sequence translation task, while *Zhang et al. (2023)*²⁴ designed a GAN to synthesize contrast-enhanced breast MRI from a combination of encoded T1-weighted MRI and DWI images.^{24,48,49,53}

However, recent promising approaches have not been used to compute synthetic subtraction images and have not been validated on their potential to improve tumor segmentation using synthetic data. Furthermore, these previous approaches generated images from a single post-contrast sequence rather than generating images from multiple temporal DCE-MRI sequences. The latter remains a largely underexplored research problem. To this end, *Schreiter et al. (2024)*⁷³ tested the simultaneous generation of DCE-MRI timepoints with a U-Net architecture using multiple inputs (T1-weighted, T2-weighted, and diffusion-weighted MRI with multiple b-values). In *Osuala et al. (2024)*,²⁶ the generation of a single, temporally variable DCE-MRI timepoint was shown by conditioning a latent diffusion model on the time-passed since respective pre-contrast acquisition. However, jointly synthesizing DCE-MRI timepoint can be desirable to ensure coherence across the resulting images. While extending over U-Net architectures by adopting a multi-scale conditional GAN, we further note that some modalities are not readily available in all clinical settings

(e.g., high-risk population DCE-MRI breast cancer screening without T2-weighted and diffusion-weighted MRI), prompting us to input only single T1-weighted MRI images into our model for contrast synthesis.

By addressing the temporal dynamics in DCE-MRI generation, we enable a more nuanced radiologic analysis of tumor localization and contrast enhancement patterns required in clinical settings. For instance, we show that our approach is promising to achieve higher image quality and detection sensitivity. At the same time, it also enables the assessment of contrast kinetics, which comprises important clinical biomarkers for cancer characterization such as contrast wash-in and wash-out slopes, peak enhancement, and perfusion and permeability parameters.^{2,8-11}

3 Materials and Methods

3.1 Dataset

The Duke-Breast-Cancer-MRI Dataset,⁴⁷ a single-institutional open-access dataset available on The Cancer Imaging Archive,⁷⁴ is used in this study. The dataset was acquired at the Duke Hospital in the United States between 1st January 2000 and 23rd March 2014. The dataset spans 922 biopsy-confirmed patient cases with invasive breast cancer. It contains information about their histology reports, demographics, treatment records, recurrence and follow-up information, ultrasound and mammography screening information, alongside a set of pre-operative MRI images. Each case involves one fat-saturated T1-weighted sequence (pre-contrast) and up to 4 corresponding fat-saturated T1-weighted DCE sequences (post-contrast). Between each DCE acquisition, a median of 131 seconds passed with scans acquired using a field strength of either 1.5 T or 3 T MRI machine. Out of the 922 patients, 828 were administered either Magnevist® or MultiHance® as CA with a contrast bolus volume ranging from 6mL to 20 mL. The axial MRI scans come in dimensions of either 320², 448² or 512² in the coronal and sagittal planes, while consisting of a varying number of slices in the axial plane. After transforming the respective DICOM files into 3D Nifti volumes, their voxel values are min-max normalized per volume and scaled to values in the range [0, 255]. Next, axial slices are extracted from the 3D fat-saturated T1-weighted (DCE-)MRI volumes after resampling them to an isotropic resolution of 1 mm³ using the pixel spacing information provided in the DICOM headers.

We further source 3D tumor segmentation masks for 254 of our cases from *Caballo et al (2023)*.² *Caballo et al (2023)* segmented these masks automatically using a fuzzy means algorithm in MATLAB. The masks were then refined by an experienced medical physicist and validated by a radiologist. We further manually verified these 254 cases, validating that the masks correctly correspond to the tumor volumes in the first phase of the DCE-MRI acquisition (timepoint 1) and, where necessary, adjusted the orientation.

For the single-sequence pre- to post-contrast synthesis model, we used 668 cases without segmentation masks, out of the total of 922 cases of the dataset, as training data, while the remaining 254 cases with masks are randomly split into validation (224 cases) and test (30 cases) sets. This split is due to the availability of ground truth tumor masks and avoids data leakage by training the synthesis model exclusively on cases that are not used for validation or test of the segmentation model. All axial slices, i.e., tumor containing and non-tumor containing slices, were extracted from 3D fat-saturated T1-weighted (DCE-)MRI Nifti volumes. These slices are then used as train, validation, and test data to enable the model to translate any 2D slice from the 3D volumes. For the

segmentation model, the same test set is used. For training and validation, we exclude 33 multi-focal cases from the 224 cases where ground truth tumor masks are available. Multi-focal cases are defined by containing multiple distinct tumor foci. Their removal allows to ensure consistency with the ground truth annotations which delineate only a single primary tumor lesion. We apply a 5-fold cross-validation, splitting the remaining 191 cases into training (80%) and validation (20%) subsets.

For the multi-sequence pre to post-contrast synthesis model, we apply a new random split of the dataset in order to have a higher number of training cases, resulting in 762 train, 50 val, and 100 test cases, and including only cases with a minimum of 3 available DCE-MRI sequences. In this multi-sequence scenario, all tumor-containing axial slices were extracted and used alongside an additional 10% of axial slices adjacent to the tumor (i.e., 5% before the first and 5% after the last tumor-containing slice in axial dimension). Tumor containing slices are identified based on the bounding box annotations of the Duke Dataset. Pre- and post-contrast slices are extracted as corresponding pairs.

3.2 Image Synthesis

3.2.1 Generative Adversarial Networks

GANs²⁷ are a family of deep generative models composed of multi-hidden layer neural networks to implicitly learn a real data distribution from a set of real data samples to then, ultimately, sample unobserved new data points from that distribution. GANs are based on a two-player min-max game of a generator and a discriminator network. The generator (G) strives to create samples ($\hat{x}$) from a noise distribution (p_z) that the discriminator (D) cannot distinguish from samples (x) stemming from the real image distribution (p_{data}), resulting in the value function:

$$\min_G \max_D V(D, G) = \min_G \max_D [\mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]]. \quad (1)$$

*Goodfellow et al. (2014)*²⁷ define the discriminator D as a binary classifier, detecting whether a sample x is either real or generated. The discriminator is trained via binary-cross entropy with the objective of minimizing the adversarial loss function $\mathcal{L}_{adv}$, which the generator, on the other hand, tries to maximise:

$$\mathcal{L}_{adv} = -\mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]. \quad (2)$$

3.2.2 Pre- to Post-Contrast DCE-MRI Synthesis

In the context of image-to-image (I2I) translation, instead of generating data from a noise distribution, GANs²⁷ receive an input sample from a source distribution (x) to generate a corresponding output sample from a target distribution ($\hat{y}$). In this research, we implement a Pix2PixHD⁷⁰ GAN for translating pre-contrast to post-contrast images. Pix2PixHD was chosen for its proven effectiveness in producing high-quality cancer imaging data,⁴ along with its network architecture and methodological approach specifically tailored for paired image-to-image translation, fitting the pre-to-post-contrast transformation scenario. As illustrated in Figure 2, the GAN architecture comprises a generator network that processes images at two different scales—one to ensure global consistency and the other to generate fine details. Additionally, the model incorporates two identical discriminator networks, each operating at different image scales based on downsampled

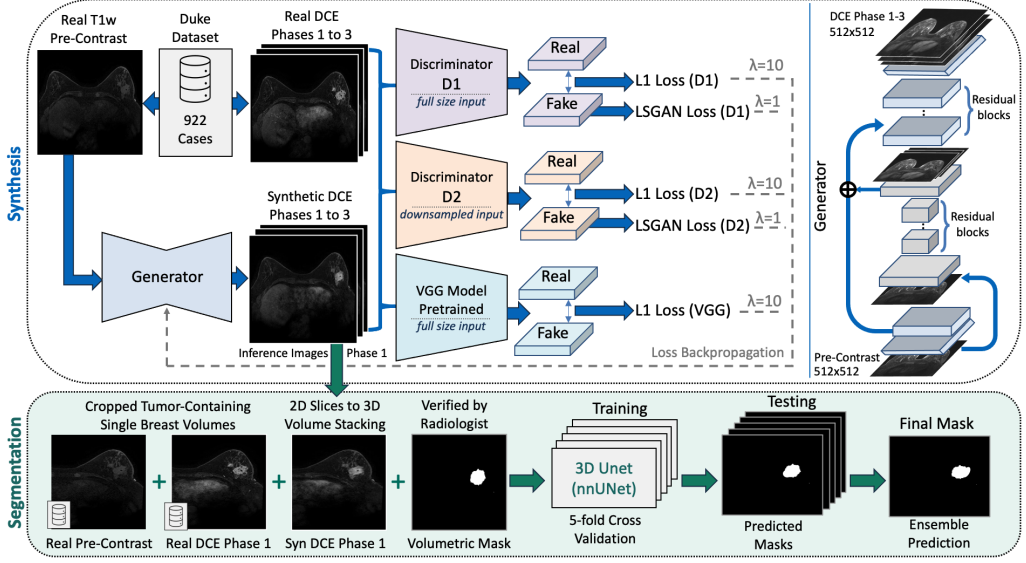


Fig 2: Overview of training workflow of our pre- to post-contrast translating GAN based on Pix2PixHD.⁷⁰ Three reconstruction losses (L1) and two least squares adversarial losses (LSGAN)⁷⁵ from two discriminators (D1 & D2) and one pre-trained VGG⁷⁶ model are backpropagated into the generator, where lambda (λ) represents the weight of each of the different losses. Processing the images at two different scales inside the generator architecture balances local detail and global consistency,⁷⁰ which is further enforced by the two different image input scales in D1 (full size) and D2 (downsampled). The segmentation method is based on 3D U-Nets⁷⁷ from the nnU-Net⁷⁸ framework. The iteratively translated synthetic post-contrast axial slices are stacked to create 3D breast MRI volumes. These synthetic volumes correspond to the tumor segmentation masks, which were initially acquired based on the real post-contrast fat-saturated sequence.

versions of the input images. The training of the model involves a weighted combination (λ) of least squares adversarial losses⁷⁵ ($\lambda_{adv} = 1$), discriminator feature matching losses ($\lambda_{fm} = 10$) calculated as the summed L1-loss between the real and synthetic image features extracted by each of the two discriminators, and a VGG-based⁷⁶ perceptual loss ($\lambda_{per} = 10$):

$$\mathcal{L}_{GAN} = \lambda_{adv} \cdot (\mathcal{L}_{adv(D1)} + \mathcal{L}_{adv(D2)}) + \lambda_{fm} \cdot (\mathcal{L}_{fm(D1)} + \mathcal{L}_{fm(D2)}) + \lambda_{per} \cdot \mathcal{L}_{per}. \quad (3)$$

Input images are transformed into the range $[-1, 1]$ and have a probability of 50% of being rotated by 90 degrees during training. Following the best practices from Wang *et al* (2018)⁷⁰, the model is trained for 200 epochs using an Adam optimizer ($\beta = 0.5$)⁷⁹ and a learning rate of $2e-4$ that decays linearly to zero from epoch 100 to 200. The images in the dataset are resized to pixel dimensions of 512×512 .

In the case of (i) pre-contrast to phase 1 DCE-MRI image synthesis, the grayscale post-contrast (phase 1) image is duplicated three times and stacked into 3 channels. This model was trained on

a single NVIDIA GeForce RTX 3090 GPU with 24GB RAM using a batch size of 1.

For the case of (ii) pre-contrast to multi-DCE phase synthesis, the respective images of DCE phase 1, 2, and 3 are concatenated resulting in an output pixel dimension of $512 \times 512 \times 3$. In this latter case, and despite outputting a single image, the pix2pixHD learns to synthesize the first three DCE-MRI acquisitions jointly. We extract the output of each of the three channel dimensions from these output images and store them separately as $512 \times 512 \times 1$ image per DCE phase. We specifically selected the first three DCE-MRI phases, as additional acquisitions, such as DCE phase 4 and phase 5, are not available for a considerable fraction of cases in the dataset. This model was trained on a single Nvidia RTX A6000 GPU with 48GB RAM and a batch size of 8 using the PyTorch library⁸⁰ in a Python 3.11 environment.

3.3 Synthetic Data Evaluation

3.3.1 Defining Trustworthy Synthetic Data

Our review of the medical image synthesis literature^{4,24,34,48,49,51–53} described in Section 2 reveals a lack of agreement on the appropriate metrics for assessing synthetic imaging data. Different metrics provide particular strengths such as correlation with human visual perception or usefulness for clinical application. In order to prioritize and select specific metrics, we note the need for a principled definition of what desirable trustworthy synthetic data should encompass. To this end, inspired by the SynTRUST framework,⁴ while also building upon insights from previous synthetic data evaluation studies and guidance provided in the literature,^{4,25,33,81–85} we provide a general definition of trustworthy synthetic data. As depicted in Fig. 3, we define trustworthy synthetic data as multi-dimensional framework comprising synthetic data (a) fidelity, (b) diversity, (c) condition adherence, (d) utility, (e) privacy, and (f) fairness.

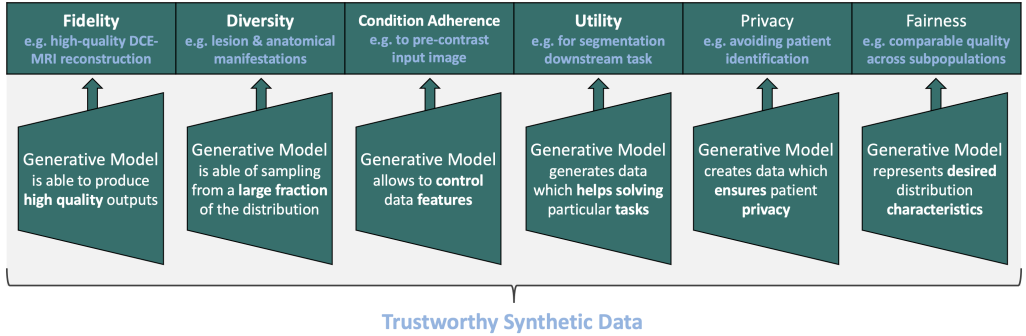


Fig 3: Depiction of the generally applicable dimensions of trustworthy synthetic data alongside respective examples (in blue font) for their adoption in deep generative models for medical image synthesis. The present study evaluates fidelity, diversity, condition adherence, and utility. Privacy and fairness are included herein for the comprehensiveness of the dimensions encompassing trustworthy synthetic data.

Fidelity (a) refers to the quality, realism, and degree to which the generated images accurately and convincingly replicate the characteristics of real-world images. Beyond visual similarity, fidelity also requires preserving essential features, shapes, textures, and statistical properties of the

original data. Diversity (b) represents the objective of generating a wide range of synthetic images, ideally capturing the full spectrum of variability present in real-world images. This includes variations in features such as intensities, texture, structure, and patterns, as well as domain shifts within the data such as differences in pathological and anatomical manifestations, viewpoints, and contexts. Condition adherence (c) describes the extent to which the generated images accurately reflect specific, predefined conditions or attributes that were set during the synthesis process. This ensures that the images conform to particular requirements or constraints, such as provided features, labels, variables, contextual details, or input images. Utility (d) represents the practical value and effectiveness of synthetic images, commonly tied to achieving one or multiple specific goals or applications. This includes how actionable synthetic images are in serving their intended purposes, such as training machine learning models, validating algorithms, translating data to another domain, or augmenting real-world datasets. Utility can be measured indirectly by quantifying the contribution and impact of synthetic images in a practical task, e.g., via ablation studies. Privacy (e) pertains to the ability of synthetic images to protect sensitive information to ensure confidentiality while still providing valuable data for analysis or training. On one hand, generated data is desired to maintain essential characteristics of the real-world dataset in order to serve specific applications, however, without revealing or compromising confidential or identifiable information such as for instance a patient’s ailments, identifying visual features or other related personal data. Fairness (f) involves ensuring that synthetic images represent diverse and balanced data across different groups or conditions, avoiding biases that could lead to undesirable outcomes. Examples include generating synthetic images that accurately reflect specific or specifically-balanced demographics, conditions, or scenarios without favoring or underrepresenting any particular sub-population. Hence, synthetic images are not to create or reinforce biases or inequities, thereby promoting equitable, impartially and inclusive performance on downstream applications.

3.3.2 Synthetic Data Evaluation Metrics for DCE-MRI Synthesis

As discussed above, it is desirable to provide a multi-faceted analysis of synthetic data while also automating the evaluation process to avoid laborious and costly manual human expert observer revisions.^{4,84,85} We assess our synthetic DCE-MRI using multiple metrics comparing them against their real counterparts hence measuring (a) fidelity, (b) diversity, and (c) condition adherence. To this end, we provide comparisons of real-synthetic image pairs and of real and synthetic image distributions. The latter is evaluated using the Fréchet Inception Distance (FID),⁸⁶ which computes the distance between two sets of features each extracted from one imaging datasets. These features are latent representations generated by passing each image in the dataset through a pretrained deep learning model, e.g. an Inception v3⁸⁷ pretrained on natural images from ImageNet⁸⁸ (FID_{Img}) or pretrained on radiology images from RadImageNet⁸⁹ (FID_{Rad}). Once the latent features are extracted from both synthetic and real datasets, they are each fitted to a multi-variate Gaussian X =real and Y =synthetic having means μ_X and μ_Y and covariance matrices Σ_X and Σ_Y in order to compute the Fréchet distance (FD) as $FD(X, Y) = \|\mu_X - \mu_Y\|_2^2 + \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{\frac{1}{2}})$. A recently introduced variation is the Fréchet Radiomics Distance²⁶ (FRD), which compares extracted handcrafted radiomics^{90,91} feature distributions extracted from medical images. To evaluate corresponding real-synthetic DCE-MRI image pairs, we use a comprehensive set of metrics including the mean squared error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), multi-scale (MS) structural similarity index measure (SSIM),⁹² as well as the Learned Perceptual

Image Patch Similarity (LPIPS).⁹³ Given the presence of corresponding reference images in the pre- to post-contrast translation of axial MRI slices, we average metrics across all MRI slice image pairs, reporting each metric alongside its standard deviation across image pairs. Lastly, we additionally assess the generated DCE-MRI images based on their (d) utility by measuring their impact when included as additional breast tumor segmentation model training data, as described in Section 3.4.

3.3.3 Scaled Aggregate Measure (SAmE)

As discussed in Section 2 and Section 3.3.1, there is no consensus on methods and metrics for evaluating of synthetic data in general, and in image-to-image synthesis tasks specifically. Various metrics are employed and reported, but there is ambiguity about which metric should take precedence, particularly when different metrics yield conflicting results. This issue also complicates determining the optimal stopping point for training a generative model. Overall, this evident lack of a consistent evaluation metric underscores the necessity for our proposed unified measure of synthetic data quality. Given that each metric captures different facets of truth, we suggest that the most effective way to evaluate synthetic data is through an ensemble of metrics. Therefore, we propose a *Scaled Aggregate Measure* (SAmE) that scales and combines several metrics. These metrics include the SSIM, MSE, MAE, FID_{Img} ,^{86,88} and FID_{Rad} .^{33,89} In this work, for simplicity, we define SAmE based on the aforementioned five metrics, but note that the integration of further metrics, such as LPIPS or PSNR, as well as additionally distinguishing between image-level and lesion-level metrics, can further improve the expressiveness and comprehensiveness of SAmE. The metrics in SAmE are scaled to a range $[0, 1]$ using per-metric min-max normalisation to achieve comparability and allow their combination. After scaling, smaller values correspond to increased performance for SAmE and each of its internal metrics, including SSIM which was reversed after scaling (i.e., the smaller, the better). Next, we compute SAmE as a non-weighted average between these metrics. The choice of the metrics in SAmE is flexible as we motivate researchers to adopt SAmE based on their particular image synthesis problem at hand. However, we draw attention that our choice of metrics comprises a complementary selection balancing perceptual metrics that capture global semantics of imaging features (FIDs), metrics of perceived quality of images (SSIM and FIDs) and metrics focusing on fine-grained pixel-level comparisons between generated and target images (MAE and MSE) to assess the accurateness of replication between an image pair. While FID is associated to a high sensitivity to small changes and high correspondence to human inspection,⁸⁵ the pixel level metrics in SAmE measure objective (MSE, MAE) and perceived image fidelity based on a combination of luminance, contrast, and structural information (SSIM).⁹⁴ We further combine analytical metrics (SSIM, MAE, MSE) with metrics derived from latent space feature representations of neural networks (FIDs), with the latter being further divided into domain-agnostic (FID_{Img}) and radiology domain-specific (FID_{Rad}) features to capture different dimensions of relevant information within the synthetic data. To this end, we compress complementary and mutually exclusive information present in the selected image quality metrics into a single meaningful measure and show its application for the problem of generative model training stopping criterion definition and training checkpoint selection.

3.4 Tumor Segmentation Downstream Task

The segmentation of tumors is an important clinical task used, among others, to analyze and quantify the tumor and its volume. An accurate tumor delineation can be used for surgery and radiation treatment planning as well as monitoring evaluating tumor growth or decline e.g. before, during and after neoadjuvant chemotherapy.

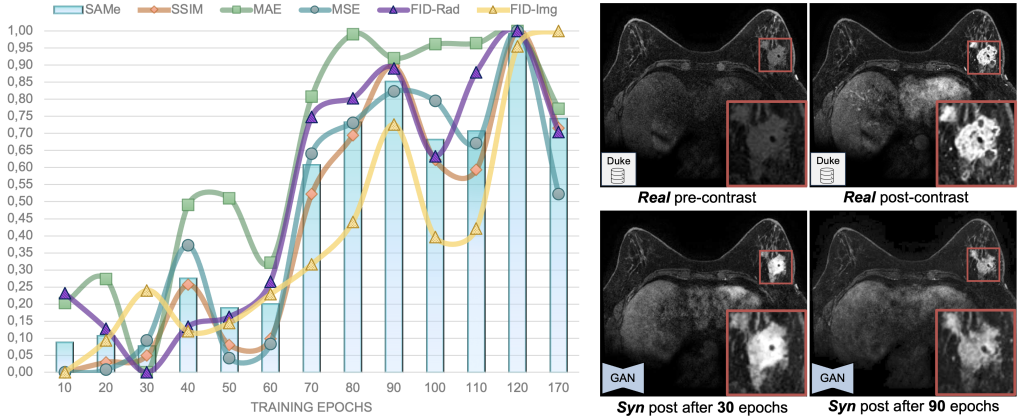
To evaluate synthetic data for automated tumor segmentation, we adopt a single 3D U-Net⁷⁷ model using the nnU-Net framework⁷⁸ (i.e., *nnunetv2 3d full_res*). nnU-Net is a versatile deep learning framework for medical image segmentation, which self-configures its architecture based on the input data. We adopt its 3D U-Net architecture to capture volumetric tumor context while also retaining fine details via its skip connections between encoding and decoding layers. Using only the 3D convolutional architecture variant of the nnU-Net framework further enables testing whether individually translated synthetic 2D breast MRI slices can be reassembled to useful 3D breast MRI volumes. We follow the vanilla nnU-Net training configuration, however, without applying any of nnU-Net’s post-processing techniques such as all-but-largest-component suppression,⁷⁸ which are not specifically targeted to our breast tumor segmentation task. We train one 3D U-Net tumor volume segmentation model for 500 epochs for each fold in a 5-fold cross-validation (CV). Test set performance is measured based on the averaged predictions of the ensemble of the five models, each of which was trained during one of the CV folds. The Dice coefficient is used as training loss and test set evaluation metric to measure segmentation performance. The Dice coefficient quantifies the overlap between the predicted tumor segmentation (A) and the ground truth (B) in a range $[0, 1]$, with 0 representing no overlap and 1 indicating complete overlap.

As shown in Fig. 2, the 2D synthetic slices are stacked to 3D synthetic volumes before being integrated as additional training data into our tumor segmentation pipeline. The same segmentation masks, which had initially been annotated in (a) the first real post-contrast DCE-MRI sequence, are used as labels in the segmentation model for (a) real post-contrast, (b) real pre-contrast, and (c) synthetic post-contrast MRI volumes, as depicted in Figure 5. Given that in the ground truth masks only the primary lesion was annotated, we remove multifocal cases (33 cases) from the segmentation dataset. We further crop the volumes to include only a single breast per image rather than both breast to avoid any bilateral cases and apply Bias field correction.⁹⁵ The segmentation models were trained on a single NVIDIA GeForce RTX 3090 GPU with 24GB RAM.

4 Experiments and Results

4.1 Selection of Generative Model

As the first experiment, we apply our proposed SAME to our single sequence DCE-MRI generative model, which translates a T1-weighted pre-contrast image to its first sequence (phase 1) DCE-MRI counterpart. To this end, we demonstrate SAME’s effectiveness as a generative model weight selection criterion. As shown in Fig. 4a, we compute SAME and its internal metrics (FID_{Img} , FID_{Rad} , MAE, MSE, SSIM) during generative model training on each 10th epoch up until epoch 170. The FID metrics are computed for 3000 and MSE, MAE and SSIM metrics for 5000 synthetic-real post-contrast axial MRI slice pairs from the validation set. In Fig. 4a, epochs with metric values close to 0 indicate good generative model performance in comparison to values close to 1. For completeness, we also provide the original values of the metrics before SAME scaling in Table 1. SAME, as the aggregate across complementary metrics, is depicted using bar charts (in blue)



(a) We inspect overall synthetic data fidelity, and diversity using the Scaled Aggregate Measure (SAME) across generative model training epochs, thereby enabling an informed selection of the best training checkpoint (i.e., epoch 30, achieving the lowest SAME). Our SAME metrics include synthetic image distribution distances (FID_{Img} and FID_{Rad}), pixel space objective (MSE and MAE) and perception-based (SSIM) quality metrics. Metrics are scaled in range $[0,1]$, where lower values indicated better performance.

(b) An illustration of synthetic breast DCE-MRI images generated during GAN training after epochs 30 and 90, exhibiting discernible variances in tumor representation. From left to right, (i) the real pre-contrast image is shown (i.e., the GAN input image), (ii) the respective real DCE phase 1 image, the synthetic real DCE phase 1 counterpart after (iii) 30 and (iv) 90 training epochs.

Fig 4: Quantitative (a) and qualitative (b) illustrations of SAME applied to DCE-MRI synthesis

in Fig. 4a. Hence, the shortest bar represents the generative model checkpoint which performs the best across metrics and epochs. We note that the generative model achieves good performance already early in training with a SAME score of 0.087 in epoch 10. The only checkpoint achieving a better results is the one in epoch 30, resulting in a SAME of 0.077. Further training follows a trend of gradually reduced performance in SAME (e.g., 0.682 in epoch 100) likely indicating an increasing overfitting on the training dataset. In the last row of Table 1 real post-contrast images are compared to their real pre-contrast counterparts. Compared to this baseline, the synthetic images from the different epochs (e.g., ep10, ep30, ep50) consistently result in overall better metrics (e.g., FID_{Img} , SSIM, MSE, albeit with the exception of MAE) when compared to real post-contrast images from the validation set.

A noteworthy observation is that, both across and within training epochs, the various metrics yield inconsistent conclusions regarding the optimal synthesis model, highlighting the need for a unified measure such as SAME. Yet, all metrics exhibit a similar overall trend, with lower (better) values until epoch 60, after which they increase remarkably, suggesting overfitting and diminishing returns from continued training. based on SAME, epoch 30 emerges as the optimal source for a model checkpoint for generating synthetic post-contrast samples, which are subsequently used for the tumor segmentation downstream task and image synthesis test set evaluation.

Table 1: Example of quantitative image quality results, based on SAmE and reported with standard deviation on the validation set where applicable, for different GAN training epochs (ep). FID_{Img} , FID_{Rad} are based on 3000 and MSE, MAE, SSIM are based on 5000 synthetic-real DCE-MRI axial MRI image slice pairs. As upper bound, the *Real DCE vs. Real Pre* baseline compares corresponding real pre-contrast and real DCE pairs.

Comparison	Dataset	Metric					
		$FID_{Img} \downarrow$	$FID_{Rad} \downarrow$	SSIM $\uparrow$	MAE $\downarrow$	MSE $\downarrow$	SAmE $\downarrow$
Real DCE vs. Syn $_{ep10}$ DCE	Val	15.047	0.108	0.701 $\pm$ 0.081	93.895 $\pm$ 41.748	37.803 $\pm$ 9.960	0.087
Real DCE vs. Syn $_{ep30}$ DCE	Val	17.308	0.081	0.699 $\pm$ 0.081	88.733 $\pm$ 39.426	38.334 $\pm$ 9.582	0.077
Real DCE vs. Syn $_{ep50}$ DCE	Val	16.412	0.089	0.696 $\pm$ 0.090	101.696 $\pm$ 44.672	38.045 $\pm$ 10.985	0.188
Real DCE vs. Syn $_{ep100}$ DCE	Val	18.778	0.219	0.669 $\pm$ 0.116	113.144 $\pm$ 59.360	42.320 $\pm$ 17.792	0.682
Real DCE vs. Real Pre	Val	34.062	0.120	0.660 $\pm$ 0.090	66.146 $\pm$ 31.758	42.933 $\pm$ 11.528	

4.2 Synthesis of First DCE-MRI Sequence

To systematically assess the quality of image synthesis, we compare metrics between synthetic and real post-contrast MRI slices in the test set. After training the generative model, we generate T1-weighted DCE-MRI phase 1 images - often corresponding to peak enhancement in the studied dataset⁴⁷ - for both the image synthesis test set (30 cases) and validation set (224 cases). Figures 5 present qualitative results, illustrating the model’s translation of entire axial breast MRI slices to the post-contrast domain, along with corresponding subtraction images for six different patient cases. It is observed that some false-positive contrast regions are hallucinated (e.g., see the 4th row), and some tumors are only partially contrast-enhanced (e.g., see the 5th row of Figure 5). In the randomly chosen patient case 045, depicted in the 6th row of Figure 5, the real post-contrast image displays hypointense areas within the tumor, suggesting the presence of a necrotic core. With this feature being not clearly visible in the pre-contrast domain, it is not reproduced in the synthetic post-contrast image. Nevertheless, the respective synthetic subtraction image enables detection and localization of the tumor, thereby preserving clinical utility for diagnostic and treatment workflows^{96,97} despite limitations in replicating the internal tumor microenvironment in full detail for contrast kinetics assessment. Overall, the qualitative outcomes of our model underscore its capability to proficiently translate pre-contrast to DCE-MRI, demonstrating strong potential in synthetic contrast localization and enhancement.

Table 2 presents a comparison of the 2D full axial slice image dataset, evaluating the similarity between the synthetic and real test case images. In this analysis, the synthetic DCE-MRI images demonstrate a significantly closer semantic and perceptual resemblance (as measured by FID scores, LPIPS, SSIM, MS-SSIM) to the real DCE-MRI phase 1 images compared to their real pre-contrast counterparts. In the comparison of *splitted test* datasets, it is important to note that the compared sets do not correspond to the same patient cases, enabling the assessment of variability across different patient test cases. Interestingly, based on the domain-agnostic FID_{Img} , the variability between real DCE-MRI cases is found to be higher than that between real and synthetic DCE-MRI cases. Conversely, the FID_{Rad} indicates less variability between real DCE-MRI datasets than between real and synthetic ones for the same dataset split. Specifically, according to the radiology domain-specific FID_{Rad} , the variability across patient cases (*splitted test*) is generally higher than the variability between the pre-, post-, and synthetic post-contrast sequences (*test*) of corresponding cases. For real vs. synthetic DCE-MRI, this also holds for FID_{Img} .

Table 2: Multi-metric synthetic image quality evaluation on the test set containing 5186 images. Synthetic images were generated after 30 GAN training epochs with SAME score as epoch selection criterion. FID_{Img} and FID_{Rad} results are based on 2000 synthetic-real phase 1 DCE-MRI axial slice pairs, while 5000 pairs were used for the remainder of the metrics. *Real DCE vs. Real Pre* describes the upper bound baseline that compares paired real pre-contrast and real phase 1 DCE slices. *Subt* refers to subtraction images, where pre-contrast images are subtracted from either their real (*Real Subt*) or synthetic (*Syn Subt*) DCE counterparts. *Splitted Test* describes a random by-patient split of the test set (i.e. without corresponding image pairs) that allows to capture the variance across patient cases in distribution comparison metrics (i.e., FID).

Comparison	Dataset	Metric							
		$FID_{Img} \downarrow$	FID_{Rad}	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	MAE $\downarrow$	MSE $\downarrow$
Real DCE vs. Syn_{ep30} DCE	Test	28.717	0.0385	0.064$\pm$0.04	32.91$\pm$1.35	0.726$\pm$0.089	0.798$\pm$0.08	85.623 $\pm$ 38.297	34.882$\pm$10.520
Real DCE vs. Real Pre	Test	59.644	0.1556	0.084 $\pm$ 0.05	32.42 $\pm$ 1.68	0.705 $\pm$ 0.104	0.780 $\pm$ 0.07	66.121$\pm$34.473	40.124 $\pm$ 16.183
Real Subt vs. Syn_{ep30} Subt	Test	46.931	0.2864	0.062 $\pm$ 0.03	34.74 $\pm$ 1.73	0.692 $\pm$ 0.097	0.717 $\pm$ 0.09	44.896 $\pm$ 23.403	23.425 $\pm$ 8.602
Real DCE vs. Syn_{ep30} DCE	Splitted Test	43.865							0.7012
Real DCE vs. Real DCE	Splitted Test	49.808	0.2060						

Additionally, we assess subtraction images created by subtracting a pre-contrast image from either its real or synthetic DCE-MRI counterpart. Compared to the real vs. synthetic DCE-MRI images, the corresponding real (*Real Subt*) vs. synthetic (*Syn_{ep30} Subt*) subtraction images show improved metrics in pretrained neural network-based image-level comparisons (i.e., LPIPS) and reconstruction-based metrics (i.e., MSE, MAE). However, at least this latter improvement can be attributed to the clipping of pixel values to 0 when they become negative after subtraction. In terms of structural perceptual metrics (i.e., SSIM, MS-SSIM) and latent feature distribution-based metrics (i.e., FID_{Rad} , FID_{Img}), the comparison between real and synthetic DCE-MRI images yields better quantitative results than the subtraction image comparison.

4.3 Tumor Segmentation in First DCE-MRI Sequence

Given the potential variability in data availability between pre- and post-contrast domains across different clinical environments, we conduct four types of tumor segmentation experiments. The first set of experiments, shown in the 1st block of Table 3, assumes that ground truth post-contrast data is entirely unavailable for segmentation during both training and testing. In this scenario, available pre-contrast training cases (*baseline 1*) are augmented with their synthetic post-contrast counterparts. The second set of experiments assumes that pre-contrast data is available for training, while the test data consists of post-contrast images. Here, we evaluate tumor segmentation performance under a domain shift, examining the impact of adding synthetic post-contrast cases to the pre-contrast *baseline 2*. The third set in Table 3 considers a scenario where real post-contrast data is available and used for both training and testing, assessing whether synthetic data can enhance the performance of the post-contrast *baseline 3*. Finally, the fourth set of experiments investigates a situation where segmentation models are trained on real post-contrast data but tested on pre-contrast cases. This scenario includes instances where contrast agents are not administered, such as in patient sub-populations like pregnant women, patients with kidney issues, those who decline contrast media, or those at high risk of allergic reactions to contrast agents.

In all data augmentation experiments, each training case is supplemented with its corresponding augmented version (e.g., real and/or synthetic post-contrast volumes). Importantly, the model

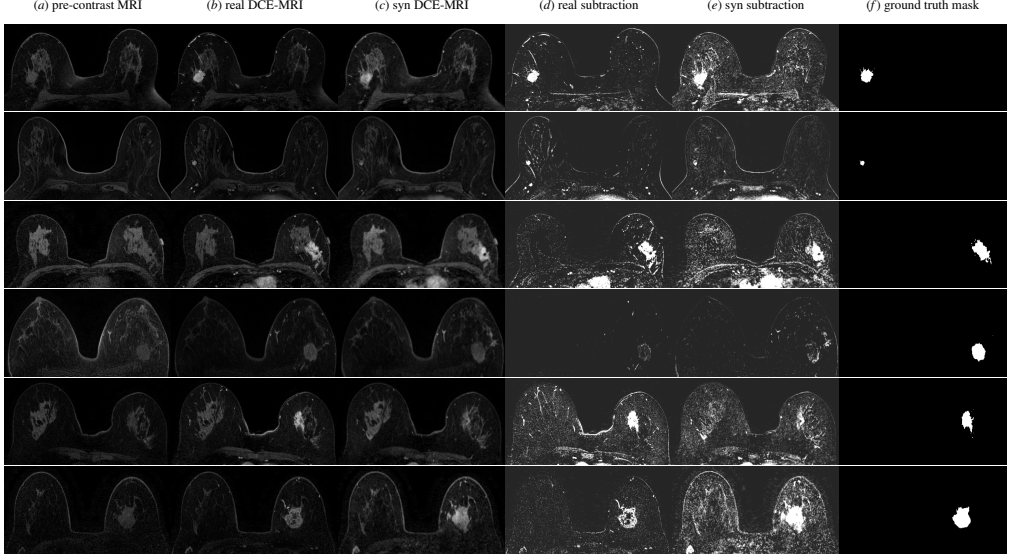


Fig 5: Synthesis of breast DCE-MRI as shown for six cases.⁴⁷ Two cases were manually selected from the validation set (1st row: Case 228 , 2nd row: Case 886), two manually selected from the test set (3rd row: Case 378, 4th row: Case 907), and two randomly selected from the test set (5th row: Case 041, 6th row: Case 045). From left to right, we illustrate axial slices of the (a) real T1-weighted pre-contrast MRI, (b) the real DCE-MRI sequence 1, (c) the synthetic DCE-MRI sequence 1, the subtraction image based on the (d) real and (e) synthetic DCE-MRI subtractions, and (f) the ground truth segmentation mask.

does not receive any indication that an original training data point (e.g., a pre-contrast volume) and its augmented counterpart (e.g., a synthetic post-contrast volume) pertain to the same patient. The reported Dice coefficients are based on ensemble predictions from five segmentation models trained using 5-fold cross-validation,⁷⁸ which is why standard deviations are not reported.

Reviewing the results for *baseline 1* in Table 3 , we observe that synthetic post-contrast augmentations do not enhance segmentation performance within the pre-contrast domain. However, in the domain shift context of *baseline 2*, the inclusion of synthetic DCE-MRI volumes leads to a marked improvement in the post-contrast domain. Specifically, augmenting real pre-contrast data with synthetic post-contrast images increases the post-contrast Dice coefficient by 0.179 (from 0.484 to 0.663), while maintaining a similar performance level in the pre-contrast domain (with Dice scores of 0.531 compared to 0.569). This finding aligns with the image quality analysis in Table 2, which confirms that the GAN-generated images fall within the post-contrast domain distribution, highlighting their effectiveness in addressing domain shift scenarios. *Baseline 3* demonstrates strong tumor segmentation performance in the post-contrast domain, achieving a Dice score of 0.790. Although the improvement with synthetic DCE-MRI augmentation is modest, it still enhances performance to 0.797, making it preferable over pre-contrast augmentations, which yield a slightly lower score of 0.780. In contrast, *baseline 4* shows a more significant Dice score increase of 0.245 (from 0.164 to 0.409) in the pre-contrast test domain when synthetic post-contrast

Table 3: Tumor volume segmentation results across four scenarios: (1) pre-contrast domain only, (2) domain shift with post-contrast testing and no access to real post-contrast data during training, (3) combined pre- and post-contrast training with post-contrast testing, and (4) synthetic post-contrast data aiding models trained in the post-contrast domain but tested on pre-contrast data (e.g., due to patient pregnancy or allergy). Synthetic data improves performance notably in domain shift and pre-contrast test cases. Reported Dice coefficients are from a model ensemble with each model trained via one 5-fold cross-validation fold.

Scenario 1. Pre-contrast training data and pre-contrast test data available		Dice $\uparrow$
<i>train on:</i>		<i>test on:</i> Real pre-contrast
Real pre-contrast (<i>baseline 1</i>)		0.569
Real pre-contrast + syn post-contrast (augmentation)		0.531
Syn post-contrast		0.486
Scenario 2. Domain shift: Pre-contrast training data, but no pre-contrast test data available		
<i>train on:</i>		<i>test on:</i> Real post-contrast
Real pre-contrast (<i>baseline 2</i>)		0.484
Real pre-contrast + syn post-contrast (augmentation)		0.663
Syn post-contrast		0.687
Scenario 3. Post-contrast training data and post-contrast test data available		Dice $\uparrow$
<i>train on:</i>		<i>test on:</i> Real post-contrast
Real post-contrast (<i>baseline 3</i>)		0.790
Real post-contrast + syn post-contrast (augmentation)		0.797
Real post-contrast + real pre-contrast (augmentation)		0.780
Real post-contrast + real pre-contrast + syn post-contrast (augmentation)		0.770
Syn post-contrast		0.687
Scenario 4. Domain shift: Post-contrast training data, but no post-contrast test data available		
<i>train on:</i>		<i>test on:</i> Real pre-contrast
Real post-contrast (<i>baseline 4</i>)		0.164
Real post-contrast + syn post-contrast (augmentation)		0.409
Syn post-contrast		0.486

augmentations are used. Despite the synthetic DCE-MRI images being closely aligned with the DCE-MRI distribution (as indicated by an FID_{Rad} of 0.0385 between synthetic and real post-contrast test data in Table 2), it nonetheless captures relevant pre-contrast signals that enable the post-contrast segmentation model to generalize more effectively to pre-contrast test data. Notably, training solely on synthetic images, without real post-contrast counterparts, further boosts segmentation performance in the post-contrast domain by 0.077 (from 0.409 to 0.486).

4.4 Joint Synthesis of Multiple DCE-MRI Sequences

Based on the respective pre-contrast T1-weighted image of each patient, we jointly generate the images corresponding to the first three DCE-MRI sequences acquisitions using a checkpoint after 30 training epochs of our multi-sequence conditional GAN.

As shown in Table 4, we use multiple metrics to quantitatively assess each generated DCE-MRI sequence based on its similarity to its respective real DCE-MRI sequence. To facilitate interpreting and evaluating the obtained metrics, we compute the metrics also for the similarity of pre-contrast images with images from each of the real DCE-MRI sequences. It is observable that across each of the three DCE-MRI sequences, the synthetically generated images by our GAN method are

Table 4: Quantitative synthetic image quality assessment for joint synthesis of multiple DCE-MRI image timepoints (Phase P1 to P3). Results include standard deviation where applicable and are based on 100 test cases consisting of 4422 image pairs. Best results in bold. *Real Post vs. Real Pre* indicates the lower bound baseline comparison between corresponding real pre-contrast MRI and DCE-MRI images. GAN denotes our proposed Pix2PixHD-based post-contrast image synthesis method. GAN Subt refers to subtraction images created by subtracting real pre-contrast images from GAN-generated DCE-MRI images. U-Net refers to the subtraction image synthesis baseline method adapted and modified from *Schreiter et al.*,⁷³ which directly predicts subtraction images from pre-contrast images. In FRD_{BB+Tex} , *BB* denotes that feature extraction is conditioned by tumor bounding box annotations,⁴⁷ while *Tex* denotes that only texture features were extracted,⁹¹ i.e., glcm, glrlm, glldm, glszm, ngtdm.

512x512 breast MRI slices with tumor		Metrics								
Set 1	Set 2	FID _{Img} ↓	FID _{Rad} ↓	FRD _{BB+Tex} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MS-SSIM ↓	MSE ↓	MAE ↓
Real Pre-Contrast	Real DCE _{P1}	39.47	0.143	209.70	.201 ± .098	31.88 ± 1.303	.712 ± .080	.785 ± .062	38.98 ± 11.88	109.55 ± 38.78
GAN: Syn DCE _{P1}	Real DCE _{P1}	20.32	14.91	116.37	.139 ± .071	32.40 ± 1.153	.749 ± .059	.862 ± .042	32.88 ± 8.331	77.65 ± 37.50
U-Net: Syn Subt _{P1}	Real Subt _{P1}	296.65	14.91	2046.91	.223 ± .713	10.67 ± 5.786	.620 ± .103	.474 ± .124	12.84 ± 7.113	21.44 ± 9.704
GAN: Syn Subt _{P1}	Real Subt _{P1}	49.22	4.531	264.25	.175 ± .073	23.63 ± 1.966	.686 ± .081	.690 ± .088	24.56 ± 9.031	51.84 ± 25.86
Real Pre-Contrast	Real DCE _{P2}	36.18	0.124	202.32	.182 ± .085	31.92 ± 1.383	.718 ± .082	.811 ± .056	38.80 ± 13.02	112.61 ± 41.55
GAN: Syn DCE _{P2}	Real DCE _{P2}	15.45	0.062	98.19	.131 ± .065	32.59 ± 1.121	.764 ± .059	.871 ± .041	31.08 ± 7.686	83.37 ± 38.74
U-Net: Syn Subt _{P2}	Real Subt _{P2}	282.00	14.07	1986.08	.212 ± .069	11.56 ± 5.771	.630 ± .105	.503 ± .131	12.58 ± 6.729	22.22 ± 10.02
GAN: Syn Subt _{P2}	Real Subt _{P2}	27.88	2.839	202.48	.156 ± .066	24.15 ± 1.99	.698 ± .079	.712 ± .091	24.25 ± 8.646	56.81 ± 25.00
Real Pre-Contrast	Real DCE _{P3}	33.79	0.121	223.68	.175 ± .082	31.95 ± 1.383	.721 ± .081	.821 ± .053	38.59 ± 12.95	112.67 ± 42.04
GAN: Syn DCE _{P3}	Real DCE _{P3}	14.38	0.079	108.13	.129 ± .066	32.64 ± 1.122	.765 ± .060	.871 ± .043	30.75 ± 7.495	80.88 ± 37.31
U-Net: Syn Subt _{P3}	Real Subt _{P3}	260.21	8.17	1679.39	.182 ± .062	12.91 ± 5.761	.638 ± .113	.550 ± .133	12.60 ± 5.684	23.82 ± 11.64
GAN: Syn Subt _{P3}	Real Subt _{P3}	49.71	0.342	155.16	.159 ± .067	23.71 ± 2.067	.683 ± .083	.692 ± .094	25.04 ± 8.563	57.41 ± 23.98

substantially closer to the real DCE-MRI images compared to the lower bound pre-contrast based baseline. This holds true across both image-distribution comparison metrics (FID_{Img}, FID_{Rad}) and image-level comparison metrics (e.g., LPIPS, SSIM, MSE). With the exception of MAE and FID_{Rad} (DCE_{P2} to DCE_{P3}), an improvement across metrics is observed with temporally later DCE-MRI acquisitions, indicating enhanced performance at subsequent imaging phases. The achieved FID_{Img} values (20.32 in DCE_{P1}) are notably low even when compared to the single DCE-MRI sequence experiments (28.71) from section 4.2 indicating that the training and joint generation using multiple DCE-MRI sequences is likely to positively affect model performance.

To quantitatively evaluate the texture of the tumor area on the synthesized images, we employ the Fréchet Radiomics Distance (FRD),²⁶ comparing radiomics texture feature distributions extracted from both the real and synthetic images. The FRD_{BB+Tex} metric we compute restricts feature extraction to tumor regions defined by bounding box annotations⁴⁷ for the 2737 respective image-annotation pairs for each dataset. In FRD_{BB+Tex} we include exclusively the texture-based features of interest, namely glcm, glrlm, glldm, glszm, and ngtdm derived using the PyRadiomics toolkit.⁹¹ The values of these features are z-score normalized across all features from both datasets (e.g. synthetic and real) and, for better interpretation, are scaled to the range common for FID as defined and recommended in *Osuala et al.*²⁶ Our adapted formulation emphasizes tumor-specific texture, which has been identified as a particularly challenging aspect of contrast-enhanced image synthesis due to the subtle and heterogeneous appearance of cancerous lesions and the tumor microenvironment. By focusing the analysis on this dimension, FRD_{BB+Tex} provides a robust measure of how well the generative model captures clinically relevant tumor characteristics in synthetic images. For our GAN-generated synthetic DCE-MRI data we observe a substantially lower

FRD_{BB+Tex} scores compared to the real pre-contrast vs real DCE comparison baseline, e.g. 116.37 and 209.70, respectively, for the first DCE-MRI timepoint. This trend is present across all of the temporal DCE-MRI sequences, demonstrating that tumor areas within our GAN-generated images overall capture a range of meaningful texture features that are present in the real DCE-MRI tumor areas. The GAN-generated subtraction images, while resulting in higher FRD_{BB+Tex} compared to its generated DCE-MRI counterparts, obtained vastly lower FRD_{BB+Tex} compared to the U-Net baseline across all DCE sequences.

Inspired by the work of Schreiter et al. (2024),⁷³ which to our knowledge represents the closest approach to multi-sequence DCE breast MRI slice generation, we implemented a benchmark U-Net model⁷⁷ for comparative evaluation. While our implementation is based on the MCO-Net architecture proposed by Schreiter et al., it has been adapted to fit the specific requirements of our application. Different to their setup, we use a batch size of 16 and apply a sigmoid activation function in the final layer instead of tanh, as our output values lie within the range [0,1]. Unlike the original work, which incorporates T2-weighted and multi-b-value diffusion-weighted images as additional inputs, we restrict our input to T1-weighted non-contrast-enhanced MRI due to the constraints of our dataset. The U-Net is trained to generate subtraction images, i.e., computed by subtracting the non-contrast T1-weighted input from each corresponding post-contrast DCE sequence, across all three DCE timepoints. While we also experimented with directly predicting post-contrast images using this U-Net approach, the subtraction-based output yielded better qualitative and quantitative performance and we report the latter in Table 4.

Overall, the U-Net generated subtraction images (compared to their real subtraction image counterparts) show considerably lower performance (apart from MSE and MAE metrics) across all DCE-MRI timepoints in terms of image distribution metrics, as well as image-to-image comparison metrics, when compared to the GAN-based approaches. The GAN-based approaches consist of (a) the comparison of GAN-generated DCE-MRI images with their real counterparts and (b) the comparison of real subtraction images with synthetic subtraction, where the latter are computed based on real pre-contrast images subtracted from GAN-generated DCE-MRI images. Subtraction images, particularly those generated by the U-Net, show strong performance on pixel-wise metrics such as MAE and MSE. This is in part likely due to the high proportion of low-intensity pixels present in both real and synthetic subtraction images, especially when compared to full DCE-MRI slices. While the GAN-generated subtraction images otherwise have generally slightly lower performance values compared to the GAN-generated DCE-MRI images, they (i) capture the real subtraction image distribution reasonably well (e.g. see FID_{Img} of 27.88 for sequence 2 in Table 4) and (ii) also show desirable results on the image-to-image comparison metrics (e.g. see LPIPS across all temporal sequences in Table 4).

Fig. 6 provides a respective qualitative comparison of three patient cases across the first three DCE-MRI sequences. In these cases, and while noting a high similarity of images across DCE-MRI acquisitions, both in the real as well as in the synthetic images a trend of increased lesion contrast enhancement towards later DCE-MRI acquisitions is noticeable. This temporal enhancement pattern is further reflected in the contrast enhancement curves, which show a generally increasing mean pixel intensity across the DCE-MRI timepoints. This trend is consistently observed in both the synthetic and real DCE data, indicating that the generative model captures patterns of the underlying contrast kinetics.

To further explore the differences between DCE-MRI acquisitions, we additionally compute the mean and standard deviation of the pixel intensity within the tumor area, which we locate based

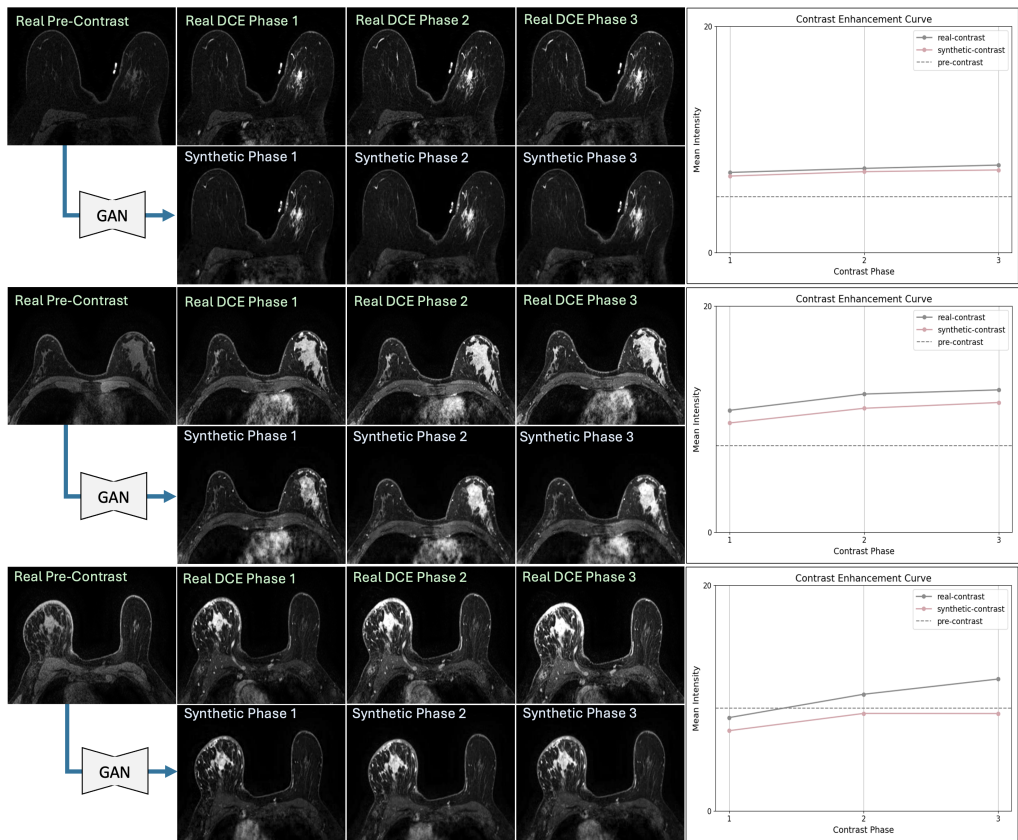


Fig 6: Qualitative results of joint multi-sequence DCE-MRI generation shown for 3 test cases with real images in each first row and respective synthetic images displayed in each second row. A single conditional GAN was trained to translate all test cases from pre-contrast to all three DCE-MRI sequences. Contrast enhancement curves are also shown, visualizing changes in image intensity across consecutive temporal DCE-MRI sequences for both real and synthetic data.

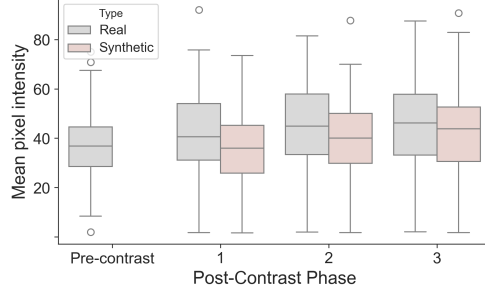


Fig 7: Illustration of temporal contrast enhancement patterns based on mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices of all test cases. While gray coloring represents values computed for pre-contrast (T1) and post-contrast (DCE) of real test case images, red denotes values extracted from their synthetic DCE-MRI counterparts.

on the bounding box information provided in the dataset.⁴⁷ The bounding box allows to capture and assess contrast uptake within the lesion but also, as opposed to lesion-level delineation, its closely surrounding tissue adjacent to the gross tumor region. Next, we aggregate the mean tumor intensity over all test cases to get a value and standard deviation for intensity for each of the three analyzed temporal DCE-MRI sequences. This process is repeated for pre-contrast and also for each synthetic DCE-MRI sequence, with the results being summarized and plotted in Fig. 7.

Focusing on the differences between real and synthetic post-contrast mean intensity temporal patterns, it is noted that the intensities generally increase towards later-stages of DCE-MRI acquisition, which is in line with aforescribed visual assessment of Fig. 6. This trend in the real mean pixel intensity patterns is also followed by their synthetic counterparts. While the synthetic intensities have a slightly lower mean value than the real DCE-MRI ones, both synthetic and real intensities have comparable corresponding variances. In both cases, these variances are larger than the one of the pre-contrast tumor area intensity. This indicates more variety in the DCE-MRI domain, which is present in the synthetic DCE-MRI images. A rationale for this diversity is given, for instance, by the inter- and intra-tumor heterogeneity often manifesting in a mixture of both hyper-intense and hypo-intense areas within the tumor in the DCE-MRI domain. Moving from generative model evaluation towards clinical application, where contrast kinetics are used as biomarkers for tumor characterization, malignancy estimation, and treatment planning, we assess intensity changes per individual tumor area across DCE-MRI sequences. To this end, we randomly select 33 test cases and visualize their mean lesion intensity value for synthetic (denoted as x markers) and real (denoted as circle markers) for each sequence as well as for the respective pre-contrast image (denoted as gray circle marker) in Fig. 8.

The observations in Fig. 8 corroborate conclusions drawn from Fig. 7 indicating the general ability of the model to capture the dynamics of contrast uptake in the tumor lesion area. In the more detailed per-case scheme in Figure 8, the majority of the cases (22 out of 33), the synthetically generated images feature the right order for the mean pixel intensity values, increasing over time. Moreover, the standard deviation value between pixel intensity values in the tumor lesion bounding box, depicted by the marker size, can also be seen to be captured overall correctly by the generative model. However, while synthetic data generally follows the trend of the real data, an

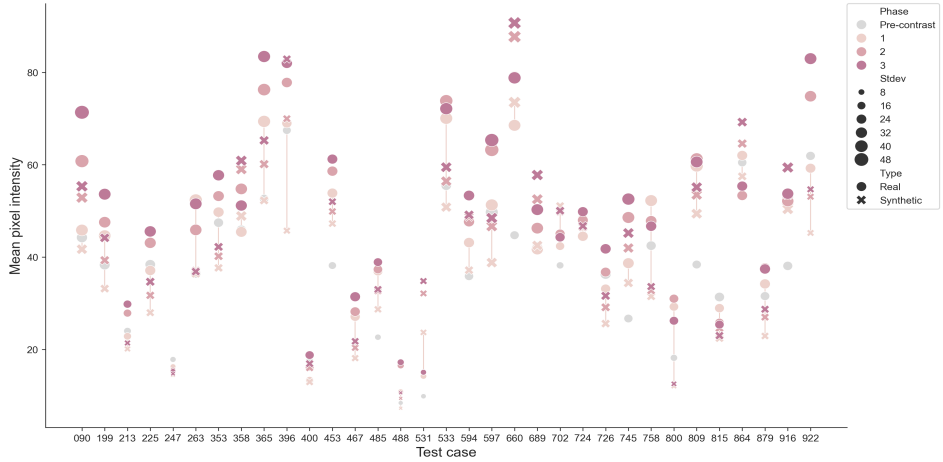


Fig 8: Visualization of temporal contrast enhancement patterns based on mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices for each of a randomly selected 33 (out of 100) test cases. Circles denote real images while x denotes synthetic counterparts. The mean pixel intensity standard deviation is represented by the marker size. The marker color encodes the temporal DCE-MRI sequence, with gray circles indicating the pre-contrast MRI sequence for comprehensiveness and comparability.

off-set between the mean intensity values for the real and the synthetic images can be observed. To this end, Fig. 8 demonstrates both, the complexity of the general task at hand, which asserts the generative model to detect, localize, highlight accordingly, and temporally adjust the highlighting for each heterogeneous tumor of each patient with variations in manufacturer, scanner, tumor molecular subtypes and patient characteristics. Overall, it is further indicated that the generative model is able to learn and represent this complexity reasonably well despite the differences between cases in contrast enhancement and tumor manifestation.

5 Discussion and Future Work

Based on the results described in Section 4, we confirm our initial hypothesis that (post-contrast) DCE-MRI imaging data can be effectively generated from pre-contrast MRI inputs. Thus, the pre-contrast MRI provides a signal that is learnable via extracting statistical patterns from patient cohorts that allow the generative model to detect, localize, and enhance cancerous lesions. Our modeling approach can be viewed as unsupervised lesion detection method that enables to identify cancerous lesions without the need of respective lesion-level annotations. Our modeling paradigm is thus of particular interest for clinical settings considering the need for cost-effective deep-learning methods, where annotation conducted by clinical experts commonly represents a major cost-driving factor. Towards clinical application it is to be further noted that synthetic data can occasionally produce false-positive tumor hallucinations. Despite of that, it can still be a valuable tool for localizing and highlighting potential anomalies within the MRI volume, as such anomalies can then be flagged for clinical revision and in-depth examination by clinical experts. In cases where imaging of both virtual contrast and injected contrast are available, the difference

it is to be explored whether insights can be derived from the difference between the contrast enhancement in both images. For instance, if the virtual contrast prediction model is trained on only cancerous lesions (or e.g., tumors of a specific molecular subtype), then a different tumor manifestation in the real contrast enhancement can indicate an out-of-distribution case (e.g., benign, different molecular subtype). This consideration sheds light on the potential lying in research synthesizing multiple possible (contrast-enhanced) lesion manifestations to explore the within-distribution diversity and, thus, the uncertainty of a contrast prediction for a specific patient case. A contrast prediction marked as being uncertain (e.g. based on a variance threshold) can warrant the administration of contrast media in a clinical setting, where the risk-benefit tradeoff of physical CA injection is evaluated on a case-by-case basis.

Apart from treatment applications, future clinical validation studies are also encouraged to investigate generative modeling methods as diagnostic tools in DCE-MRI as screening modality in high-risk populations (e.g., with change in BRCA1 or BRCA2). Validating CA-free MRI screening regimes with synthetic DCE-MRI requires evaluation on additional dataset, where, apart from cancerous lesions, a variety of benign findings are present. As compared to the present dataset assembled from neoadjuvant treatment cohorts, tumors in the screening regime can differ being smaller in size, earlier in stage, while also the potentially younger median age of patients can affect manifestations on the imaging data.

Going beyond unsupervised lesion detection, the prediction of contrast enhancement kinetics via multi-sequence DCE-MRI generation unlocks further additional clinical use-cases. As also indicated by our work, the progress in the field of computer vision, which rapidly expands the capabilities of deep generative models, narrows the gap towards wide-spread clinical application of such complex temporal modeling tasks. We note that multi-sequence DCE-MRI synthesis can potentially result in higher quality synthetic images than single-sequence synthesis, likely due to the generative model having to learn more nuanced patterns in the temporal synthesis task. Considering our exploratory results for multi-sequence synthesis of DCE-MRI, such synthetic temporal contrast enhancement patterns can be considered a promising alternative for patients where CA injection is too risky. In this realm, biomarkers based on contrast kinetics are to be further investigated, particularly towards a comparison between real and synthetic biomarkers. This can enable future studies to define a benchmark for the research community that assesses the clinical meaningfulness of a generative model’s produced synthetic biomarkers. Such a benchmark can further be extended by assessing the usefulness of real and synthetic biomarkers for specific clinical downstream tasks such as lesion malignancy prediction or treatment response prognosis. We note additional potential for enhancement of our approach by integrating time-since-event variables²⁶ that condition the generative model to generate DCE-MRI images for specific moments in time e.g. based on milliseconds passed since pre-contrast acquisition or CA injection. This methodological extension introduces new possibilities of clinical application such as synthetically covering time-points of interest in DCE-MRI protocols that could not have been acquired, for instance due to the limitation of lengthy acquisition times of MRI scans. Multi-variable conditioning of the generative model can further allow counterfactual generation of DCE-MRI sequences, where for instance the malignancy level, tumor subtype, patient age, tumor staging, selected treatment, or other variables can be varied thereby resulting in insightful comparisons of alternative tumor manifestations.

While an exhaustive hyperparameter search was beyond the scope of this work due to computational constraints, future research is to explore alternative training schedules and optimizer configurations,^{98–100} enhancing the stability and generalization of our generative models. Notably,

the observed performance degradation beyond epoch 30 suggests that more refined learning rate strategies can help mitigate overfitting and enable sustained training stability. In this context, a future direction is to quantify the impact of these strategies, such as cosine annealing learning rate schedules, which have been shown to facilitate smoother convergence and help avoid sharp minima¹⁰¹ and learning rate warm-up strategies.¹⁰²

Additional potential realms of research include the comparison of fine-tuning vs training from scratch of the generative model, assessing different GAN and generative model architectures such as denoising probabilistic diffusion models.^{28–30} Using larger-scale datasets will further invite for a stratified analysis as per tumor type, CA type, bolus volume, scanner type, field strength, and clinical center. A further line of research is to investigate the effect of integrating additional 2D and 3D imaging data such as diffusion-weighted (DWI) MRI, T2-weighted MRI, non-fat-saturated images as well as subtraction images either as (additional) inputs or as outputs that can be useful to enhance image fidelity.^{24,73,103} In this regard, such input modalities provide complementary information to conditioning the generative model likely resulting in increased anatomical accuracy and reliability of lesion characteristics of the synthetic post-contrast sequences, particularly in complex cases. For instance, T2-weighted images can depict lesion morphology better as well as perifocal or prepectoral edema within the breast, while DWI captures higher signal intensity without relying on contrast agent administration.⁹⁶ However, including these modalities is associated with additional challenges, such as the need for accurate image registration across modalities to avoid the introduction of respective alignment errors or artifacts, as well as the limited availability of such MRI acquisitions in real-world settings.

Another avenue to explore is to probe whether DCE-MRI synthesis quality can be enhanced by (a) iteratively enlarging the proportion of tumor-containing slices or (b) by gradually increasing the patch size around the tumor region-of-interest¹⁰⁴ during training.

As exemplified by row 1 in Figure 5, we observe the difficulty in distinguishing fibroglandular tissue from tumor enhancement in contrast enhancement synthesis. This distinction is inherently challenging, especially in dense breasts, where benign fibroglandular enhancement can mimic malignant uptake. To avoid such cases to lead to an increase in false positives future work can apply region-aware loss functions or incorporating annotated fibroglandular maps can improve tissue differentiation in complex breast compositions. In general, adding and correctly weighting a reconstruction, perceptual, or adversarial loss for only the tumor area in DCE-MRI synthesis holds potential to increase fidelity, diversity, and usefulness of the synthetic data. Such approaches can potentially allow models to better capture particular those tumor features that lie in the long-tail of the distribution such as, for instance, tumors with particular necrotic cores. Future studies can include these tumor-focused modifications additionally in the evaluation methods extending SAME and its components by adding and weighting tumor-area restrict computation of metrics for 2D slices as well as 3D volumes.

Based on tumor segmentation, we show that stacked synthetic DCE-MRI volumes can be useful in increasing the robustness of downstream task models. Enhancing segmentation model generalizability across imaging domains is particularly advantageous in DCE-MRI, especially for patient populations restricted to pre-contrast imaging due to CA administration contraindications such as allergic reactions, absence of consent, pregnancy, or compromised renal function. Additionally, in some cases, fat-saturated DCE-MRI sequence may closely mimic pre-contrast images due to small-sized tumors, low CA doses or rapid washout, emphasizing the necessity for models that maintain robust performance across both pre-contrast and post-contrast settings.

Focusing on the use of synthetic DCE-MRI for training via augmentation and domain adaptation, another dimension worth exploring is testing models trained on real post-contrast images using synthetic post-contrast images at inference. While this may offer technical insights into the interchangeability of real and synthetic data, it remains a clinically open and critical question whether decision-making can rely on synthetic inputs, given the risk of artifacts or hallucinations.^{4,105,106} The clearer clinical utility of synthetic data lies in enhancing model performance as training input, both within and across modalities, where it has shown value in improving generalization and robustness.^{33,44,107,108}

As an alternative to data augmentation, we guide future work to also explore pre-training on synthetic DCE-MRI for downstream model training. Related to that, a promising approach is to analyze the impact of concatenated corresponding multi-modality image inputs into the segmentation model during training and testing (e.g., pre-contrast with multiple DCE-MRI sequences, as well as synthetic sequences) while also separately evaluating challenging bilateral and multifocal cases. While this study explores synthetic DCE-MRI in the context of tumor segmentation, additional downstream applications such as radiomic feature analysis represent important directions for future research. Radiomic features quantify tumor heterogeneity, morphology, and enhancement dynamics, which are essential for diagnosis and treatment planning.⁹⁰ In this context, the proposed SAME metric may serve as a useful proxy for evaluating the preservation of structural and textural features relevant to radiomic tasks. Conversely, insights from radiomic analysis can inform future extensions of the SAME metric to enhance its interpretability and robustness as a quantification tool of medical image synthesis. As a unified score of multiple quantitative measures, a next step in the evaluation of SAME is to assess its alignment with clinical relevance. The latter can be assessed based on radiologist feedback or expert-driven image validation to correlate SAME scores with visual assessments, diagnostic confidence, or lesion detectability. Guiding potential refinements to its weighting or formulation, such future work can help to establish the metric in clinical workflows.

6 Conclusion

Following the SynTRUST framework⁴ for trustworthy medical image synthesis studies, our work demonstrates that virtual contrast injection can generate high-quality synthetic DCE-MRI images, effectively supporting tumor detection, localization, and segmentation. Our findings highlight the potential of integrating deep generative models into MRI workflows as a non-invasive alternative for patients who cannot undergo standard contrast-enhanced imaging, thereby enabling more accurate and personalized treatment strategies. The potential of DCE-MRI synthesis is further demonstrated as a training data augmentation method to enhance downstream breast tumor segmentation models, which, for instance, increases robustness across modalities. We further define trustworthy synthetic data based upon which we introduce the SAME as a unified metric to evaluate synthetic data quality and guide generative model training checkpoint selection, addressing the limitations of conventional single-metric assessments. Additionally, generating multiple subsequent DCE-MRI sequences facilitates a deeper assessment of tumor response to contrast media, offering critical insights for tumor characterization and individualized care planning. Overall, this work marks a significant step toward incorporating virtual contrast and deep generative models into clinical practice, paving the way for improved diagnostic accuracy and patient outcomes in breast cancer management.

Disclosures

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Acknowledgements

This project has received funding from the European Union’s Horizon Europe and Horizon 2020 research and innovation programme under grant agreement No 101057699 (RadioVal) and No 952103 (EuCanImage), respectively. Also, this work was partially supported by the project FUTURE-ES (PID2021-126724OB-I00) and AIMED (PID2023-146786OB-I00) from the Ministry of Science, Innovation and Universities of Spain. We would like to thank Dr Marco Caballo (Radboud University Medical Center, The Netherlands) for providing 254 segmentation masks for the Duke Dataset. RO acknowledges a research stay grant from the Helmholtz Information and Data Science Academy (HIDA). Daniel M. Lang and Julia A. Schnabel received funding from HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator.

Data, Materials, and Code Availability

The data⁴⁷ used in this study is publicly available on <https://doi.org/10.7937/TCIA.e3sv-re93>. The code is available in a Github repository on <https://github.com/RichardObi/SimulatingDCE>. The trained models are readily usable in the *medigan* library on <https://github.com/RichardObi/medigan>.

References

- 1 Global Cancer Observatory, “The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research.” <https://gco.iarc.fr/> (2023). Accessed: 2023-08-07.
- 2 M. Caballo, W. B. Sanderink, L. Han, *et al.*, “Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI,” *Journal of Magnetic Resonance Imaging* **57**(1), 97–110 (2023).
- 3 S. Joshi, R. Osuala, L. Garrucho, *et al.*, “Leveraging epistemic uncertainty to improve tumour segmentation in breast mri: an exploratory analysis,” in *Medical Imaging 2024: Image Processing*, **12926**, 292–300, SPIE (2024).
- 4 R. Osuala, K. Kushibar, L. Garrucho, *et al.*, “Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging,” *Medical Image Analysis*, 102704 (2022).
- 5 D. M. Lang, E. Schwartz, C. I. Bercea, *et al.*, “Multispectral 3d masked autoencoders for anomaly detection in non-contrast enhanced breast mri,” in *MICCAI Workshop on Cancer Prevention through Early Detection*, 55–67, Springer (2023).
- 6 S. Radhakrishna, S. Agarwal, P. M. Parikh, *et al.*, “Role of magnetic resonance imaging in breast cancer management,” *South Asian journal of cancer* **7**(02), 069–071 (2018).
- 7 M. C. Comes, D. La Forgia, V. Didonna, *et al.*, “Early prediction of breast cancer recurrence for patients treated with neoadjuvant chemotherapy: a transfer learning approach on dce-mris,” *Cancers* **13**(10), 2298 (2021).
- 8 S. Wu, W. A. Berg, M. L. Zuley, *et al.*, “Breast mri contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer,” *Breast Cancer Research* **18**, 1–10 (2016).

- 9 R. H. El Khouli, K. J. Macura, M. A. Jacobs, *et al.*, “Dynamic contrast-enhanced mri of the breast: quantitative method for kinetic curve type assessment,” *American Journal of Roentgenology* **193**(4), W295–W300 (2009).
- 10 C. K. Kuhl, P. Mielcareck, S. Klaschik, *et al.*, “Dynamic breast mr imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions?,” *Radiology* **211**(1), 101–110 (1999).
- 11 X. Tao, L. Wang, Z. Hui, *et al.*, “Dce-mri perfusion and permeability parameters as predictors of tumor response to ccrt in patients with locally advanced nslc,” *Scientific reports* **6**(1), 35569 (2016).
- 12 Y.-C. C. Chang, E. Ackerstaff, Y. Tschudi, *et al.*, “Delineation of tumor habitats based on dynamic contrast enhanced mri,” *Scientific reports* **7**(1), 9746 (2017).
- 13 V. A. Arasu, R. C. Chen, D. N. Newitt, *et al.*, “Can signal enhancement ratio (ser) reduce the number of recommended biopsies without affecting cancer yield in occult mri-detected lesions?,” *Academic radiology* **18**(6), 716–721 (2011).
- 14 P. Marckmann, L. Skov, K. Rossen, *et al.*, “Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging,” *Journal of the American Society of Nephrology* **17**(9), 2359–2362 (2006).
- 15 C. Olchoway, K. Cebulski, M. Łasecki, *et al.*, “The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-a systematic review,” *PloS one* **12**(2), e0171704 (2017).
- 16 J.-M. Idée, M. Port, I. Raynal, *et al.*, “Clinical and biological consequences of transmetalation induced by contrast agents for magnetic resonance imaging: a review,” *Fundamental & clinical pharmacology* **20**(6), 563–576 (2006).
- 17 T. Kanda, K. Ishii, H. Kawaguchi, *et al.*, “High signal intensity in the dentate nucleus and globus pallidus on unenhanced t1-weighted mr images: relationship with increasing cumulative dose of a gadolinium-based contrast material,” *Radiology* **270**(3), 834–841 (2014).
- 18 N. C. Nguyen, T. T. Molnar, L. G. Cummin, *et al.*, “Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine administrations: a retrospective analysis,” *Radiology* **296**(1), 122–130 (2020).
- 19 European Medicines Agency (EMA), “EMA’s final opinion confirms restrictions on use of linear gadolinium agents in body scans.” <https://www.ema.europa.eu/> (2023). Online; accessed 06 August 2023.
- 20 R. J. McDonald, J. C. Weinreb, and M. S. Davenport, “Symptoms associated with gadolinium exposure (sage): a suggested term,” *Radiology* **302**(2), 270–273 (2022).
- 21 J. Rogowska, E. Olkowska, W. Ratajczyk, *et al.*, “Gadolinium as a new emerging contaminant of aquatic environments,” *Environmental toxicology and chemistry* **37**(6), 1523–1534 (2018).
- 22 H. M. Dekker, G. J. Stroomberg, A. J. Van der Molen, *et al.*, “Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents,” *Insights into Imaging* **15**(1), 1–14 (2024).
- 23 R. Brünjes and T. Hofmann, “Anthropogenic gadolinium in freshwater and drinking water systems,” *Water research* **182**, 115966 (2020).

- 24 T. Zhang, L. Han, A. D’Angelo, *et al.*, “Synthesis of Contrast-Enhanced Breast MRI Using Multi-b-Value DWI-based Hierarchical Fusion Network with Attention Mechanism,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 79–88, Springer Nature Switzerland, (Cham) (2023).
- 25 R. Osuala, S. Joshi, A. Tsirikoglou, *et al.*, “Pre-to post-contrast breast mri synthesis for enhanced tumour segmentation,” in *Medical Imaging 2024: Image Processing*, **12926**, 226–237, SPIE (2024).
- 26 R. Osuala, D. Lang, P. Verma, *et al.*, “Towards learning contrast kinetics with multi-condition latent diffusion models,” *arXiv preprint arXiv:2403.13890* (2024).
- 27 I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, 2672–2680 (2014).
- 28 J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, *et al.*, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*, 2256–2265, PMLR (2015).
- 29 Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in Neural Information Processing Systems* **32** (2019).
- 30 J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020).
- 31 R. Rombach, A. Blattmann, D. Lorenz, *et al.*, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695 (2022).
- 32 W. H. Pinaya, M. S. Graham, E. Kerfoot, *et al.*, “Generative AI for Medical Imaging: extending the MONAI Framework,” *arXiv preprint arXiv:2307.15208* (2023).
- 33 R. Osuala, G. Skorupko, N. Lazrak, *et al.*, “medigan: a Python library of pretrained generative models for medical image synthesis,” *Journal of Medical Imaging* **10**(6), 061403–061403 (2023).
- 34 L. Pasquini, A. Napolitano, M. Pignatelli, *et al.*, “Synthetic post-contrast imaging through artificial intelligence: clinical applications of virtual and augmented contrast media,” *Pharmaceutics* **14**(11), 2378 (2022).
- 35 F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, *et al.*, “Denoising diffusion probabilistic models for 3D medical image generation,” *Scientific Reports* **13**(1), 7303 (2023).
- 36 P. Isola, J.-Y. Zhu, T. Zhou, *et al.*, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134 (2017).
- 37 J.-Y. Zhu, T. Park, P. Isola, *et al.*, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).
- 38 Y. Choi, M. Choi, M. Kim, *et al.*, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797 (2018).
- 39 T. Park, M.-Y. Liu, T.-C. Wang, *et al.*, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346 (2019).

- 40 V. Sushko, E. Schönfeld, D. Zhang, *et al.*, “You only need adversarial supervision for semantic image synthesis,” *arXiv preprint arXiv:2012.04781* (2020).
- 41 C. Saharia, W. Chan, H. Chang, *et al.*, “Palette: Image-to-image diffusion models,” in *ACM SIGGRAPH 2022 conference proceedings*, 1–10 (2022).
- 42 L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847 (2023).
- 43 J. M. Wolterink, A. M. Dinkla, M. H. Savenije, *et al.*, “Deep mr to ct synthesis using unpaired data,” in *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2*, 14–23, Springer (2017).
- 44 L. Garrucho, K. Kushibar, R. Osuala, *et al.*, “High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection,” *Frontiers in Oncology* **12**, 1044496 (2023).
- 45 E. García, X. Lladó, R. Mann, *et al.*, “Breast composition measurements from full-field digital mammograms using generative adversarial networks,” in *17th International Workshop on Breast Imaging (IWBI 2024)*, **13174**, 181–188, SPIE (2024).
- 46 N. Konz, Y. Chen, H. Dong, *et al.*, “Anatomically-controllable medical image generation with segmentation-guided diffusion models,” *arXiv preprint arXiv:2402.05210* (2024).
- 47 A. Saha, M. R. Harowicz, L. J. Grimm, *et al.*, “A machine learning approach to radio-genomics of breast cancer: a study of 922 subjects and 529 dce-mri features,” *British journal of cancer* **119**(4), 508–516 (2018).
- 48 G. Müller-Franzes, L. Huck, S. Tayebi Arasteh, *et al.*, “Using Machine Learning to Reduce the Need for Contrast Agents in Breast MRI through Synthetic Images,” *Radiology* **307**(3), e222211 (2023).
- 49 Y. Xue, B. E. Dewey, L. Zuo, *et al.*, “Bi-directional synthesis of pre-and post-contrast mri via guided feature disentanglement,” in *International Workshop on Simulation and Synthesis in Medical Imaging*, 55–65, Springer (2022).
- 50 C. Chen, C. Raymond, W. Speier, *et al.*, “Synthesizing mr image contrast enhancement using 3d high-resolution convnets,” *IEEE Transactions on Biomedical Engineering* **70**(2), 401–412 (2022).
- 51 E. Kim, H.-H. Cho, J. Kwon, *et al.*, “Tumor-attentive segmentation-guided gan for synthesizing breast contrast-enhanced mri without contrast agents,” *IEEE Journal of Translational Engineering in Health and Medicine* **11**, 32–43 (2022).
- 52 J. Zhao, D. Li, Z. Kassam, *et al.*, “Tripartite-gan: Synthesizing liver contrast-enhanced mri to improve tumor detection,” *Medical image analysis* **63**, 101667 (2020).
- 53 P. Wang, P. Nie, Y. Dang, *et al.*, “Synthesizing the first phase of dynamic sequences of breast mri for enhanced lesion identification,” *Frontiers in Oncology* **11**, 792516 (2021).
- 54 M. Hasny, O. B. Demirel, A. Amyar, *et al.*, “Myocardial scar enhancement in lge cardiac mri using localized diffusion,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 307–316, Springer (2024).
- 55 A. Bône, S. Ammari, J.-P. Lamarque, *et al.*, “Contrast-enhanced brain mri synthesis with deep learning: key input modalities and asymptotic performance,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1159–1163, IEEE (2021).

- 56 J. Liu, Y. Li, S. Gao, *et al.*, “Trustworthy contrast-enhanced brain mri synthesis,” *arXiv preprint arXiv:2407.07372* (2024).
- 57 S. U. Dar, M. Yurt, L. Karacan, *et al.*, “Image synthesis in multi-contrast mri with conditional generative adversarial networks,” *IEEE transactions on medical imaging* **38**(10), 2375–2388 (2019).
- 58 E. Gong, J. M. Pauly, M. Wintermark, *et al.*, “Deep learning enables reduced gadolinium dose for contrast-enhanced brain mri,” *Journal of magnetic resonance imaging* **48**(2), 330–340 (2018).
- 59 K.-H. Cheng, W. Li, F. K.-H. Lee, *et al.*, “Pixelwise gradient model with gan for virtual contrast enhancement in mri imaging,” *Cancers* **16**(5), 999 (2024).
- 60 A. Rofena, V. Guarrasi, M. Sarli, *et al.*, “A deep learning approach for virtual contrast enhancement in contrast enhanced spectral mammography,” *Computerized Medical Imaging and Graphics* **116**, 102398 (2024).
- 61 J. Liu, Y. Tian, C. Duzgol, *et al.*, “Virtual contrast enhancement for ct scans of abdomen and pelvis,” *Computerized Medical Imaging and Graphics* **100**, 102094 (2022).
- 62 J. Kleesiek, J. N. Morshuis, F. Isensee, *et al.*, “Can virtual contrast enhancement in brain mri replace gadolinium?: a feasibility study,” *Investigative radiology* **54**(10), 653–660 (2019).
- 63 C. J. Preetha, H. Meredig, G. Brugnara, *et al.*, “Deep-learning-based synthesis of post-contrast t1-weighted mri for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study,” *The Lancet Digital Health* **3**(12), e784–e794 (2021).
- 64 L. Gui, C. Ye, and T. Yan, “Cavm: Conditional autoregressive vision model for contrast-enhanced brain tumor mri synthesis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 161–170, Springer (2024).
- 65 M. Piening, F. Altekrieger, G. Steidl, *et al.*, “Conditional generative models for contrast-enhanced synthesis of t1w and t1 maps in brain mri,” *arXiv preprint arXiv:2410.08894* (2024).
- 66 A. Durrer, J. Wolleb, F. Bieder, *et al.*, “Diffusion models for contrast harmonization of magnetic resonance images,” *arXiv preprint arXiv:2303.08189* (2023).
- 67 M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, 214–223, PMLR (2017).
- 68 T. Pinetz, E. Kobler, R. Haase, *et al.*, “Faithful synthesis of low-dose contrast-enhanced brain mri scans using noise-preserving conditional gans,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 607–617, Springer (2023).
- 69 Y. Lei, L. Xu, X. Wang, *et al.*, “Ifgan: Pre-to post-contrast medical image synthesis based on interactive frequency gan,” *Electronics* **13**(22), 4351 (2024).
- 70 T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, *et al.*, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807 (2018).
- 71 G. Müller-Franzes, L. Huck, M. Bode, *et al.*, “Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast mri,” *European Radiology Experimental* **8**(1), 53 (2024).
- 72 L. Han, T. Tan, T. Zhang, *et al.*, “Synthesis-based imaging-differentiation representation learning for multi-sequence 3d/4d mri,” *arXiv preprint arXiv:2302.00517* (2023).

- 73 H. Schreiter, J. Eberle, L. A. Kapsner, *et al.*, “Virtual dynamic contrast enhanced breast mri using 2d u-net architectures,” in *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, 85–95, Springer (2024).
- 74 K. Clark, B. Vendt, K. Smith, *et al.*, “The cancer imaging archive (tcia): maintaining and operating a public information repository,” *Journal of digital imaging* **26**, 1045–1057 (2013).
- 75 X. Mao, Q. Li, H. Xie, *et al.*, “Least squares generative adversarial networks,” in *Proceedings of the IEEE international Conference on Computer Vision*, 2794–2802 (2017).
- 76 K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556* (2014).
- 77 O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* **18**, 234–241, Springer (2015).
- 78 F. Isensee, P. F. Jaeger, S. A. Kohl, *et al.*, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods* **18**(2), 203–211 (2021).
- 79 D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980* (2014).
- 80 A. Paszke, S. Gross, F. Massa, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems* **32**, H. Wallach, H. Larochelle, A. Beygelzimer, *et al.*, Eds., 8024–8035, Curran Associates, Inc. (2019).
- 81 R. Osuala, D. M. Lang, A. Riess, *et al.*, “Enhancing the utility of privacy-preserving cancer classification using synthetic data,” *arXiv preprint arXiv:2407.12669* (2024).
- 82 X. Xing, F. Felder, Y. Nan, *et al.*, “You don’t have to be perfect to be amazing: Unveil the utility of synthetic images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 13–22, Springer (2023).
- 83 M. Woodland, M. A. Taie, J. A. M. Silva, *et al.*, “Importance of Feature Extraction in the Calculation of Fréchet Distance for Medical Imaging,” *arXiv preprint arXiv:2311.13717* (2023).
- 84 A. Borji, “Pros and cons of gan evaluation measures,” *Computer vision and image understanding* **179**, 41–65 (2019).
- 85 A. Borji, “Pros and cons of gan evaluation measures: New developments,” *Computer Vision and Image Understanding* **215**, 103329 (2022).
- 86 M. Heusel, H. Ramsauer, T. Unterthiner, *et al.*, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” *Advances in Neural Information Processing Systems* **30** (2017).
- 87 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
- 88 J. Deng, W. Dong, R. Socher, *et al.*, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, 248–255, Ieee (2009).
- 89 X. Mei, Z. Liu, P. M. Robson, *et al.*, “Radimagenet: an open radiologic deep learning research dataset for effective transfer learning,” *Radiology: Artificial Intelligence* **4**(5), e210315 (2022).

- 90 P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, “Radiomics: extracting more information from medical images using advanced feature analysis,” *European Journal of Cancer* **48**(4), 441–446 (2012).
- 91 J. J. Van Griethuysen, A. Fedorov, C. Parmar, *et al.*, “Computational radiomics system to decode the radiographic phenotype,” *Cancer Research* **77**(21), e104–e107 (2017).
- 92 Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing* **13**(4), 600–612 (2004).
- 93 R. Zhang, P. Isola, A. A. Efros, *et al.*, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).
- 94 T. Samajdar and M. I. Quraishi, “Analysis and evaluation of image quality metrics,” in *Information Systems Design and Intelligent Applications*, J. K. Mandal, S. C. Satapathy, M. Kumar Sanyal, *et al.*, Eds., 369–378, Springer India, (New Delhi) (2015).
- 95 N. J. Tustison, B. B. Avants, P. A. Cook, *et al.*, “N4itk: improved n3 bias correction,” *IEEE transactions on medical imaging* **29**(6), 1310–1320 (2010).
- 96 R. M. Mann, N. Cho, and L. Moy, “Breast mri: state of the art,” *Radiology* **292**(3), 520–536 (2019).
- 97 D. Saslow, C. Boetes, W. Burke, *et al.*, “American cancer society guidelines for breast screening with mri as an adjunct to mammography,” *CA: a cancer journal for clinicians* **57**(2), 75–89 (2007).
- 98 I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101* (2017).
- 99 H. Hazimeh and N. Ponomareva, “Mind the (optimality) gap: A gap-aware learning rate scheduler for adversarial nets,” in *International Conference on Artificial Intelligence and Statistics*, 3018–3033, PMLR (2023).
- 100 Y. Zhang, C. Chen, Z. Li, *et al.*, “Adam-mini: Use fewer learning rates to gain more,” *arXiv preprint arXiv:2406.16793* (2024).
- 101 I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983* (2016).
- 102 D. S. Kalra and M. Barkeshli, “Why warmup the learning rate? underlying mechanisms and improvements,” *Advances in Neural Information Processing Systems* **37**, 111760–111801 (2024).
- 103 A. Liebert, H. Schreiter, L. A. Kapsner, *et al.*, “Impact of non-contrast-enhanced imaging input sequences on the generation of virtual contrast-enhanced breast mri scans using neural network,” *European radiology* , 1–14 (2024).
- 104 S. M. Fischer, L. Felsner, R. Osuala, *et al.*, “Progressive growing of patch size: Resource-efficient curriculum learning for dense prediction tasks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 510–520, Springer (2024).
- 105 J. P. Cohen, M. Luck, and S. Honari, “Distribution matching losses can hallucinate features in medical image translation,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*, 529–536, Springer (2018).

- 106 M. Tivnan, S. Yoon, Z. Chen, *et al.*, “Hallucination index: An image quality metric for generative reconstruction models,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 449–458, Springer (2024).
- 107 R. Dorent, A. Kujawa, M. Ivory, *et al.*, “Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation,” *Medical Image Analysis* **83**, 102628 (2023).
- 108 V. M. Campello, P. Gkontra, C. Izquierdo, *et al.*, “Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge,” *IEEE Transactions on Medical Imaging* **40**(12), 3543–3554 (2021).

Biographies and photographs of the other authors are not available.

List of Figures

- 3 Depiction of the generally applicable dimensions of trustworthy synthetic data alongside respective examples (in blue font) for their adoption in deep generative models for medical image synthesis. The present study evaluates fidelity, diversity, condition adherence, and utility. Privacy and fairness are included herein for the comprehensiveness of the dimensions encompassing trustworthy synthetic data.
- 4 Quantitative (a) and qualitative (b) illustrations of SAME applied to DCE-MRI synthesis
- 5 Synthesis of breast DCE-MRI as shown for six cases.⁴⁷ Two cases were manually selected from the validation set (1st row: Case 228 , 2nd row: Case 886), two manually selected from the test set (3rd row: Case 378, 4th row: Case 907), and two randomly selected from the test set (5th row: Case 041, 6th row: Case 045). From left to right, we illustrate axial slices of the (a) real T1-weighted pre-contrast MRI, (b) the real DCE-MRI sequence 1, (c) the synthetic DCE-MRI sequence 1, the subtraction image based on the (d) real and (e) synthetic DCE-MRI subtractions, and (f) the ground truth segmentation mask.
- 6 Qualitative results of joint multi-sequence DCE-MRI generation shown for 3 test cases with real images in each first row and respective synthetic images displayed in each second row. A single conditional GAN was trained to translate all test cases from pre-contrast to all three DCE-MRI sequences. Contrast enhancement curves are also shown, visualizing changes in image intensity across consecutive temporal DCE-MRI sequences for both real and synthetic data.
- 7 Illustration of temporal contrast enhancement patterns based on mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices of all test cases. While gray coloring represents values computed for pre-contrast (T1) and post-contrast (DCE) of real test case images, red denotes values extracted from their synthetic DCE-MRI counterparts.

- 8 Visualization of temporal contrast enhancement patterns based on mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices for each of a randomly selected 33 (out of 100) test cases. Circles denote real images while x denotes synthetic counterparts. The mean pixel intensity standard deviation is represented by the marker size. The marker color encodes the temporal DCE-MRI sequence, with gray circles indicating the pre-contrast MRI sequence for comprehensiveness and comparability.

List of Tables

- 1 Example of quantitative image quality results, based on SAME and reported with standard deviation on the validation set where applicable, for different GAN training epochs (ep). FID_{Img} , FID_{Rad} are based on 3000 and MSE, MAE, SSIM are based on 5000 synthetic-real DCE-MRI axial MRI image slice pairs. As upper bound, the *Real DCE vs. Real Pre* baseline compares corresponding real pre-contrast and real DCE pairs.
- 2 Multi-metric synthetic image quality evaluation on the test set containing 5186 images. Synthetic images were generated after 30 GAN training epochs with SAME score as epoch selection criterion. FID_{Img} and FID_{Rad} results are based on 2000 synthetic-real phase 1 DCE-MRI axial slice pairs, while 5000 pairs were used for the remainder of the metrics. *Real DCE vs. Real Pre* describes the upper bound baseline that compares paired real pre-contrast and real phase 1 DCE slices. *Subt* refers to subtraction images, where pre-contrast images are subtracted from either their real (*Real Subt*) or synthetic (*Syn Subt*) DCE counterparts. *Splitted Test* describes a random by-patient split of the test set (i.e. without corresponding image pairs) that allows to capture the variance across patient cases in distribution comparison metrics (i.e., FID).
- 3 Tumor volume segmentation results across four scenarios: (1) pre-contrast domain only, (2) domain shift with post-contrast testing and no access to real post-contrast data during training, (3) combined pre- and post-contrast training with post-contrast testing, and (4) synthetic post-contrast data aiding models trained in the post-contrast domain but tested on pre-contrast data (e.g., due to patient pregnancy or allergy). Synthetic data improves performance notably in domain shift and pre-contrast test cases. Reported Dice coefficients are from a model ensemble with each model trained via one 5-fold cross-validation fold.

- 4 Quantitative synthetic image quality assessment for joint synthesis of multiple DCE-MRI image timepoints (Phase P1 to P3). Results include standard deviation where applicable and are based on 100 test cases consisting of 4422 image pairs. Best results in bold. *Real Post vs. Real Pre* indicates the lower bound baseline comparison between corresponding real pre-contrast MRI and DCE-MRI images. GAN denotes our proposed Pix2PixHD-based post-contrast image synthesis method. GAN Subt refers to subtraction images created by subtracting real pre-contrast images from GAN-generated DCE-MRI images. U-Net refers to the subtraction image synthesis baseline method adapted and modified from *Schreiter et al.*⁷³ which directly predicts subtraction images from pre-contrast images. In FRD_{BB+Tex} , *BB* denotes that feature extraction is conditioned by tumor bounding box annotations,⁴⁷ while *Tex* denotes that only texture features were extracted,⁹¹ i.e., glcm, glrlm, gldm, glszm, ngtdm.

Chapter 4

Time-Conditioned Pre- To Post-Contrast Breast MRI Synthesis using Latent Diffusion Models

This chapter presents a novel method to simulate breast MRI contrast enhancement without using gadolinium-based (GA) agents, addressing safety concerns like allergic reactions, kidney complications, and environmental impact associated with GA administration. Exploring and leveraging ControlNet [234] and Latent Diffusion Models (LDMs) [181], this approach generates synthetic DCE-MRI sequences from pre-contrast images, supporting unsupervised cancer detection while reducing reliance on physical contrast agents. Going beyond the dissertation author’s previous work [166], this study, among others, proposes a methodology that integrates the exact timepoint of each synthetic DCE-MRI image into the generation process thereby allowing to iteratively generate an image for each DCE-MRI acquisition, but also for all timepoints in between physical acquisitions. This is particularly important considering that in clinical settings the time passed between acquisition varies and depends, for instance, on scanner setup and scanning times, patient well-being and physiology especially after contrast agent injection, as well as hospital policies and number of scans to be obtained. This variation is thus integrated into our modelling approach enhancing the

flexibility of the generative model. A further novel consideration is the importance of biomarkers in the assessment of temporal contrast enhancement patterns that, for example, characterize disease type and severity. This motivates our proposal of imaging biomarkers and, in particular, their variation between real and synthetic data, as means of evaluation of synthetic DCE-MRI images in particular, as well as synthetic medical imaging data at large.

4.1 Towards learning contrast kinetics with multi-condition latent diffusion models

Summary:

This paper introduces a novel approach to synthesizing dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) sequences using diffusion models. DCE-MRI, critical for the detection and characterization of tumors, traditionally requires gadolinium-based contrast agents, which are associated to several health risks and require an invasive procedure. To address this, ContrastControlNet (CC-Net) is proposed, a diffusion model capable of generating post-contrast MRI images from non-contrast data, simulating the time-dependent contrast uptake in tumors. The model incorporates imaging metadata and time intervals to generate realistic, multi-sequence MRI images, offering a potential non-invasive alternative to contrast agents.

In addition to the diffusion model, the paper introduces the Fréchet Radiomics Distance (FRD), a novel evaluation metric that measures the quality of synthetic medical images by comparing biomarker variability between real and generated data. Through qualitative and quantitative validation on a breast DCE-MRI dataset, it is demonstrated that the model produces high-quality synthetic images that closely resemble real contrast-enhanced images, both visually and in terms of key clinical features. The FRD metric is shown to effectively capture the quality of such synthetic images, making this method a promising tool for unsupervised non-invasive tumor detection, localization and characterization, e.g., via the analysis of contrast kinetics patterns.

Main Contributions:

- **Development of a Multi-conditional Latent Diffusion Model:** The paper introduces a novel diffusion model (CC-Net) for generating contrast-enhanced MRI sequences from non-contrast images. It thus provides a potential alternative to the use of contrast agents addressing the associated health risks and expanding

the applicability of DCE-MRI to patient populations where contrast injection is contraindicated.

- **Simulation of Time-dependent Contrast Uptake:** This is the first work to use diffusion models to simulate the time-dependent contrast uptake patterns in medical imaging, particularly DCE-MRI, enabling the generation of dynamic sequences from static pre-contrast images.
- **Introduction of Fréchet Radiomics Distance (FRD):** The paper proposes and validates the FRD as a new image quality metric specifically designed for evaluating synthetic medical images by measuring biomarker variability between real and synthetic data.

Contributions | The dissertation author led the conceptualization and development of the study including the manuscript preparation as well as the ideation, design, implementation, validation, and analysis of the multi-conditional latent diffusion model, the synthesis of multiple DCE-MRI timepoints, as well as the Fréchet Radiomics Distance (FRD). Lang, D.M., Verma, P., Joshi, S., Tsirikoglou, A., Skorpuko, G., Kushibar, K., Garrucho, L., Pinaya, W.H.L., Diaz, O., Schnabel, J.A., Lekadir, K. provided critical feedback and assisted in manuscript preparation. Joshi, S., Garrucho, L., and Lang, D.M. helped in the curation of the dataset. Tsirikoglou, A. assisted in the visualization of multi-sequence DCE-MRI synthesis results. Verma, P. assisted in the validation of the Fréchet Radiomics Distance.

Declaration of Doctoral Use | None of the co-authors have used, either implicitly or explicitly, this publication for the development of their own doctoral thesis, nor do they intend to do so in the future.

Source Code | The reproducible source code of this work is made readily available at <https://github.com/Richard0bi/ccnet>. The code providing the Fréchet Radiomics Distance library is made readily available at <https://github.com/Richard0bi/frd-score>.

Python Package | The *frd-score* library is available via the pip package manager from the Python Packaging Index (PyPI) at <https://pypi.org/project/frd-score>.

Publication | Conference: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 - 27th International Conference, Proceedings, Part V, Series ISSN: 0302-9743, Publisher: Springer Cham, CORE2023: Rank A.

Copyright Statement | Reproduced with permission from Springer Nature (License Number: 5985211006609). This contribution was originally published as: "**Osuala, R.**, Lang, D.M., Verma, P., Joshi, S., Tsirikoglou, A., Skorpuko, G., Kushibar, K., Garrucho, L., Pinaya, W.H.L., Diaz, O., Schnabel, J.A., Lekadir, K. (2024). To-

4.1. Contribution: MICCAI 2024 (CCNet & FRD)

wards Learning Contrast Kinetics with Multi-condition Latent Diffusion Models. *In International Conference on Medical Image Computing and Computer-Assisted Intervention 2024. Oct 4 (pp. 713-723). Cham: Springer Nature Switzerland.*"
https://doi.org/10.1007/978-3-031-72086-4_67



Towards Learning Contrast Kinetics with Multi-condition Latent Diffusion Models

Richard Osuala^{1,2,3(✉)}, Daniel M. Lang^{2,3}, Preeti Verma¹, Smriti Joshi¹,
Apostolia Tsirikoglou⁴, Grzegorz Skorupko¹, Kaisar Kushibar¹,
Lidia Garrucho^{1,4}, Walter H. L. Pinaya⁵, Oliver Diaz^{1,6}, Julia A. Schnabel^{2,3,5},
and Karim Lekadir^{1,7}

¹ Departament de Matemàtiques i Informàtica, Universitat de Barcelona,
Barcelona, Spain
richard.osuala@ub.edu

² Helmholtz Center Munich, Munich, Germany

³ Technical University of Munich, Munich, Germany

⁴ Karolinska Institutet, Stockholm, Sweden

⁵ Kings College London, London, UK

⁶ Computer Vision Center, Bellaterra, Spain

⁷ Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract. Contrast agents in dynamic contrast enhanced magnetic resonance imaging allow to localize tumors and observe their contrast kinetics, which is essential for cancer characterization and respective treatment decision-making. However, contrast agent administration is not only associated with adverse health risks, but also restricted for patients during pregnancy, and for those with kidney malfunction, or other adverse reactions. With contrast uptake as key biomarker for lesion malignancy, cancer recurrence risk, and treatment response, it becomes pivotal to reduce the dependency on intravenous contrast agent administration. To this end, we propose a multi-conditional latent diffusion model capable of acquisition time-conditioned image synthesis of DCE-MRI temporal sequences. To evaluate medical image synthesis, we additionally propose and validate the Fréchet radiomics distance as an image quality measure based on biomarker variability between synthetic and real imaging data. Our results demonstrate our method's ability to generate realistic multi-sequence fat-saturated breast DCE-MRI and uncover the emerging potential of deep learning based contrast kinetics simulation. We publicly share our accessible codebase at <https://github.com/RichardObi/ccnet> and provide a user-friendly library for Fréchet radiomics distance calculation at <https://pypi.org/project/frd-score>.

Keywords: Contrast Agent · Synthesis · DCE-MRI · Generative Models

1 Introduction

Magnetic resonance imaging (MRI) and, in particular, dynamic contrast enhanced (DCE)-MRI are remarkably sensitive and effective modalities for tumor detection, localization and characterization and, thus, have become ubiquitous in clinical cancer treatment planning and monitoring. The uptake of contrast in DCE-MRI sequences plays a pivotal role as biomarker for cancer detection, tumor molecular subtype and malignancy differentiation, as well as cancer recurrence and treatment response prediction [1, 21]. However, intravenously-injected gadolinium-based contrast agents (CA) used in DCE-MRI have been associated with a wide range of concerns [11], such as a risk of nephrogenic systemic fibrosis, bioaccumulation in the brain, and its invasiveness causing a non-applicability in patient populations with pregnancy, adverse reactions, kidney malfunction or where consent is missing. Furthermore, due to its multiple temporal acquisitions, DCE-MRI is costly and time-consuming, prone to motion artifacts, and the contrast injection an uncomfortable procedure for patients [11, 12, 25].

Public health institutions, such as the European Medicines Agency, recommend to restrict gadolinium-based CA [3], which further emphasizes the need to develop alternative methods. A potentially faster, cost-effective, motion artifact-free, and non-invasive alternative is the synthetic generation of DCE-MRI using deep generative models. First studies applied generative adversarial networks (GANs) [4] to generate post-contrast images from pre-contrast images [10, 12, 25]. Recent works have further used diffusion models [17] and latent diffusion models (LDMs) [15] to synthesize medical images [2, 7, 14], such as pre-contrast breast MRI conditioned on anatomical segmentation masks [8]. Despite their recent successes in computer vision, diffusion models and LDMs have, to the best of our knowledge, not been applied to pre- to post-contrast DCE-MRI translation.

A largely unaddressed aspect in medical image synthesis is domain-specific image quality evaluation. To date, popular methods, such as the Fréchet inception distance (FID) [5], are based on feature extractor models trained on natural image datasets. Despite a considerable domain gap, these methods are commonly applied to medical imaging data without alteration, thereby failing to capture medical nuances such as abnormalities [2]. Recently, an FID calculation based on a radiology domain-specific feature extractor was proposed [13], which nevertheless showed some volatility and no significant correlation with human judgement [20]. A further limitation of such methods is the pretraining of the feature-extractor on 2D data (with unknown inherent biases), which is not applicable nor readily extendable to 3D medical images.

In this work, we aim to address the aforementioned gaps, resulting in the following three contributions:

- Design, implementation and validation of a multi-conditional latent diffusion model for pre- to post-contrast MRI synthesis.
- We present the first work that simulates time-dependent contrast uptake on imaging data using diffusion models.

- We propose and validate the Fréchet radiomics distance (FRD), a novel radiology-specific quality evaluation method of 3D and 2D synthetic images based on biomarker variability.

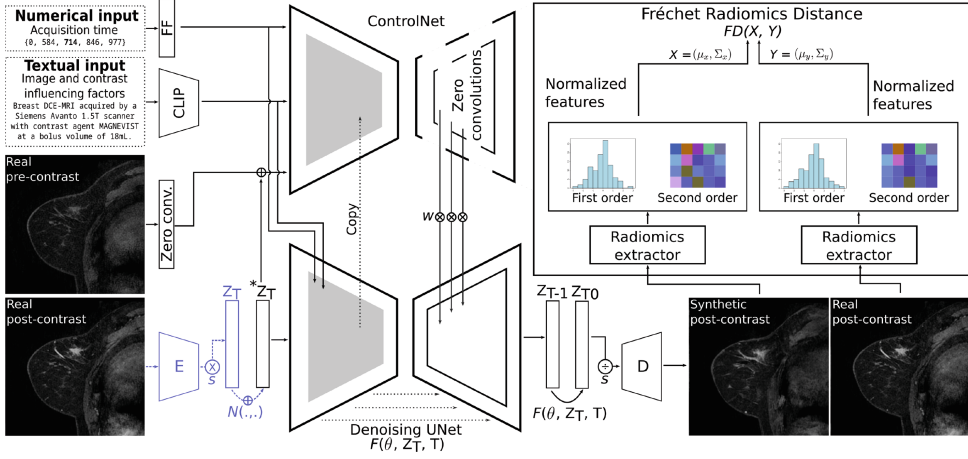


Fig. 1. Overview of our proposed methods, including ContrastControlNet (CCNet) and the Fréchet radiomics distance (FRD). CCNet trains the denoising U-Net and the ControlNet in consecutive stages under contrast enhancement-specific conditioning (pre-contrast image, text, acquisition time). During inference, E is discarded (in violet) and, based on a random latent z_T and w -weighted ControlNet guidance, the U-Net generates the post-contrast image latent z_{T0} . z_{T0} is divided by factor S and decoded via D into image space. Finally, FRD compares extracted real and synthetic imaging biomarker distributions. (Color figure online)

2 Methodology

2.1 Multi-condition Latent Diffusion Model

Recently, diffusion models [6, 17] have emerged as a promising new class of generative models due to their exceptional ability to model complex distributions. Such diffusion models consist of two processes, namely, a forward diffusion process and a reverse denoising process. The forward process is a Gaussian transition that gradually destroys the structure of a real data point $x_0 \sim p(x_0)$ by adding noise with different scales to obtain a series of noisy variables $x_1, x_2, \dots, x_T$:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}) \mathbf{I}). \quad (1)$$

The reverse process is parameterized by another Gaussian transition which gradually denoises x_T in T timesteps resulting in restoration of initial data point x_0 :

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1} | \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \quad (2)$$

Latent diffusion models (LDM) [15] are designed to iteratively denoise a learned latent representation z_T of image x_T rather than operating directly on image space, which allows for more memory-efficient training and improved conditioning. The latter includes high-quality image synthesis based on textual input [2, 15]. However, text descriptions have been shown to be either inefficient or insufficient to accurately convey detailed controls upon the final generated images (e.g., to control the fine-grained semantic layout).

ControlNet [23] introduces plug-and-play condition encoders tailored for pre-trained diffusion models. We adopt ControlNet as auxiliary encoder to integrate the conditioning pre-contrast image while preserving the integrity of the original post-contrast LDM generator. As shown in Fig. 1, the ControlNet encoder outputs are reintegrated into the diffusion model through zero-convolutional layers. We enhance this approach by extracting and propagating relevant textual metadata. We parse a clinical tabular dataset to assemble free-text prompts based on factors influencing DCE-MRI contrast manifestation. The final text is input to the denoising U-Net of both the LDM and the ControlNet via cross-attention and contains manufacturer, scanner, field strength, contrast agent, and bolus volume information. Furthermore, we integrate the time passed since pre-contrast acquisition into the model. Here, we extract pre- and post-contrast acquisition times from respective DICOM headers and input them as continuous variables into two dense layers before concatenating the resulting output with the timestep embedding of ControlNet and the denoising U-Net. Through optimization of the mean squared error between predicted and added noise per timestep, our ContrastControlNet (CC-Net) method learns to (i) extract meaningful patterns from pre-contrast, (ii) interpret conditions, (iii) reconstruct the corresponding post-contrast image, (iv) localize the lesion, and (v) predict realistic hyper- and hypo-intense lesion contrast uptake patterns.

2.2 Biomarker Variability as Image Quality Measure

Despite the significant domain gap stemming from its feature extractor being trained on natural RGB images, FID [5] is a commonly used metric in medical imaging, measuring the diversity and fidelity of synthetic images via real-synthetic distribution comparison. However, high-fidelity (and high-diversity) synthetic data can still be of low clinical utility and vice versa [22]. For analyzing synthetic data, we note the need to ideally encompass (a) image utility indicators, (b) medical imaging domain-specific measurement, (c) scalability from 2D to 3D settings, and (d) explainable results based on interpretable features. Balancing these requirements, we propose measuring synthetic data quality based on imaging biomarker variability. In particular, we measure the feature distribution distance of multiple normalized biomarker values extracted from real and synthetic images. In radiology settings, 2D and 3D radiomics features can be extracted as non-invasive imaging biomarkers that quantify phenotypic characteristics [9, 19]. Further noting the capabilities of radiomics to capture tumor heterogeneity [9], or as predictor of treatment response [1] and tumor subtype

[16] in DCE-MRI, we introduce the Fréchet radiomics distance (FRD) as synthetic data quality measure. While the selection of radiomics features included in FRD is flexible, we compute FRD based on the common set of features suggested by [19] including first-order statistics ($n = 19$), co-occurrence gray level matrix (GLM) ($n = 24$), run length GLM ($n = 16$), size zone GLM ($n = 16$), neighbouring gray tone difference matrix ($n = 5$) and dependence GLM ($n = 14$). Features are computed given the image and, optionally, a respective region of interest annotation. As depicted in Fig. 1, to compute the FRD between two imaging datasets, we extract and normalize FRD features per image and dataset and model the resulting two feature sets as Gaussian distributions, allowing us to compute a distance between them. Hence, for each image x_i in a dataset, we extract a value v_{ji} for each FRD feature j . Then, each v_{ji} is min-max normalized based on the values $v_{j1}, v_{j2}, \dots, v_{jn}$ over all images X in the dataset. Next, the resulting values v_j of feature j are scaled to the common range observed for FID latent feature values, i.e. $[0, 7.456]$. This calibration later allows for interpretation of final FRD value and its comparison to FID, considering the intuition the image synthesis field has developed for FID value interpretation [2, 13, 14, 20, 22]. The obtained synthetic and real feature sets V are fitted to multivariate Gaussian distributions. These distributions are defined by their means (μ) and covariance matrices (Σ). The FRD value is computed as the dissimilarity between real data X and synthetic data Y via the Fréchet distance defined as:

$$FD(X, Y) = \|\mu_X - \mu_Y\|_2^2 + \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{\frac{1}{2}}). \quad (3)$$

3 Experiments and Results

3.1 Dataset and Implementation

In this study, we use the public Duke-Breast-Cancer-MRI Dataset [16]. The dataset and its imaging metadata encompasses 922 biopsy-confirmed breast cancer cases, each comprising one fat-saturated T1 sequence (pre-contrast) and up to 4 corresponding fat-saturated T1-weighted DCE sequences (post-contrast) with a median of 131 s passed between DCE sequences. The 1.5T or 3T MRI scans come in dimensions of either 320^2 , 448^2 or 512^2 in the coronal and sagittal planes, with varying slice numbers in the axial plane. As we extract tumor-containing axial slices, we note considerable changes between pre- and post-contrast in non-tumor related areas (e.g., heart area). Hence, we crop the extracted slices first increasing the width and height of the tumor bounding box to half the width and height of the full image (e.g., 224^2 in case of a 448^2 image) thereby resulting in a tumor-containing single breast image. We split the dataset by patient into train ($n = 842$), validation ($n = 50$), and test ($n = 30$) sets. All models were trained on a single GPU, either Nvidia A100 (80 GB RAM) or RTX A6000 (48 GB RAM), using Python3.11's pytorch, diffusers and monai-generative libraries [14].

3.2 Fréchet Radiomics Distance as Image Perturbation Correlate

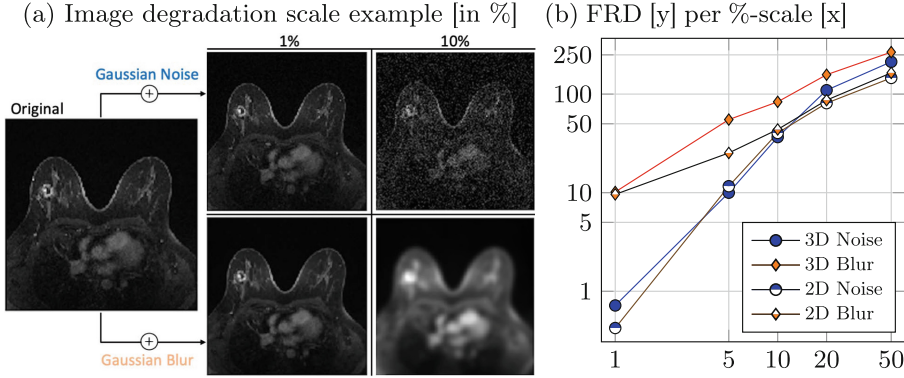


Fig. 2. (a) Image perturbation scales in breast MRI, (b) resulting Fréchet radiomics distance (FRD) values (y-axis) per percentage scale (x-axis) per applied perturbation for 2D axial slices and for 3D volumes based on DCE-MRI post-contrast phase 1 data from 254 patient cases.

Adopting the validation strategy of the FID in its original publication [5], we observe the correlation between the Fréchet radiomics distance (FRD) and the amount by which the quality of an imaging dataset is reduced. Concretely, we compare FRD feature distributions between an unchanged imaging dataset and its quality-reduced equivalent, where the quality reduction is based on a scaling factor. As visualized in Fig. 2a, we apply this procedure to 254 DCE-MRI (phase 1) cases of the Duke Dataset using Gaussian noise and Gaussian blurring for image perturbation on scales from 1% up to 50%. Unlike the FRD results shown in Table 1, in Fig. 2b we calculate the scores on full axial slices using the tumor mask. We observe FRD monotonically increasing with perturbation scale demonstrating FRD’s capability of capturing the quality-reduction level for both 2D and 3D data.

3.3 Generation of DCE Sequences from Pre-Contrast Images

In CC-Net, we first initialize a pretrained autoencoder (AE) from stable diffusion (SD), and then (b) proceed to train the denoising U-Net (i.e., LDM) before (c) training the ControlNet, after which we finally (d) run inference on the test set. In each step, we observe several hyperparameters to influence the empirical results. In (a), we note visible differences between the generally accurate breast MRI reconstructions of different pretrained SD AEs. We select the AE from SD *2-1-base* over *v1-5* and *xl-base-1.0* [15]. In (b), we notice a high dependence of output quality on scaling factor s with which the AE representation is multiplied before being used in U-Net training. We find $s = 0.1$ to improve results

upon the recommended $s = 0.18215$. We further identify a tendency for exploding gradients which was better addressed by clipping the gradient value rather than its norm (e.g., at 15). Both fine-tuning the SD U-Net, but also training it from scratch produced desirable outputs. We use a DDPM [6] noise scheduler with 1000 timesteps, AdamW as optimizer, and batch sizes of 8, 16, and 32, and learning rates of 2.5^{-5} and 5^{-5} , for which we obtain similar results. The U-Net is trained for 100 epochs and selected from the epoch with lowest validation loss for further use in (c) and (d). In (c), we follow the hyperparameter setup from (b) with half the batch size and no gradient clipping. In (d), we increase inference speed by using a denoising diffusion implicit model (DDIM) [18] scheduler with 200 timesteps without visible performance decrease. We set the text guidance scale to 1 as higher scales did not improve output quality, however, increasing the ControlNet guidance weight from 1 to a value in range (1, 1.6] enhanced perceived image quality. In (a)–(c), following AE pretraining, the 224×224 input images are stacked in 3-channels and normalized in range $[-1, 1]$ before applying small-scale intensity and affine augmentations. As observable in Fig. 3a, the contrast-enhanced tumors account for a large part of the difference between pre- and post-contrast images. Bearing this in mind, we design experiments comparing the image and distribution-wise difference between our synthetic and the real post-contrast images against the difference between corresponding real pre-contrast and real post-contrast images. To quantify this difference, we use our FRD and the common FID [5], LPIPS [24] and mean squared error (MSE) metrics. Obtained results alongside ablations are summarized in Table 1. Overall, CC-Net achieves substantially better results than the pre-contrast baseline across metrics and across post-contrast phases. Surprisingly, providing textual information not necessarily increases the performance of CC-Net_{Any}. Training the LDM from scratch instead of SD fine-tuning rather decreases performance (e.g., FID). The impact of longer ControlNet training is positive although marginal. Increasing the ControlNet guidance weight can be beneficial, and, in line with qualitative visual analysis, improves FRD and LPIPS. In Fig. 3b, using CC-Net_{Any}, we further analyze the pixel intensity distributions in the tumor area across phases aggregated over test cases. Synthetic images noticeably follow the principal trend of real contrast kinetics, despite a scale difference. A similar pattern is observable when taking contrast enhancements outside the tumor area into account by normalizing (dividing) mean tumor intensity by mean intensity of all other (tumor-free) pixels.

Table 1. Synthetic image quality evaluation based on FRD, FID, LPIPS, and MSE metrics. Where applicable, results are reported with standard deviation and are based on 30 test cases consisting of 1010 images in each of the post-contrast phases P1, P2, and P3, and 417 images in phase P4. *Any* refers to time-conditioned model training on all post-contrast phases. In *+TXT* textual input is used in training and inference, *+LDM* refers to LDM model training from scratch as opposed to stable diffusion 2-1-base fine-tuning. *+CG* refers to a ControlNet guidance weight increase, e.g. from 1 to 1.6, during inference. *+LT* stands for a 50 epoch longer training of ControlNet. Best results are in bold.

224 × 224 Single Breasts with Tumor		Metrics			
Set 1	Set 2	FRD ↓	FID ↓	LPIPS ↓	MSE ↓
Real Pre-Contrast	Real DCE-P1	49.07	68.20	0.223 ± .102	51.18 ± 17.80
CC-Net _{Any} P1	Real DCE-P1	35.64	41.38	0.192 ± .070	46.92 ± 15.40
CC-Net _{Any+Txt} P1	Real DCE-P1	39.98	42.85	0.186 ± .073	47.20 ± 15.76
CC-Net _{Any+Txt+LDM} P1	Real DCE-P1	41.55	62.41	0.200 ± .074	49.17 ± 14.48
CC-Net _{Any+Txt+LDM+CG1.6} P1	Real DCE-P1	22.50	64.61	0.194 ± .072	46.12 ± 14.21
CC-Net _{Any+Txt+LDM+LT} P1	Real DCE-P1	45.19	60.58	0.193 ± .073	48.08 ± 14.44
CC-Net _{Any+Txt+LDM+CG1.6+LT} P1	Real DCE-P1	37.70	62.21	0.192 ± .072	45.48 ± 13.60
Real Pre-Contrast	Real DCE-P2	74.26	64.90	0.212 ± .095	50.00 ± 16.85
CC-Net _{Any} P2	Real DCE-P2	58.07	40.36	0.191 ± .076	46.10 ± 14.19
Real Pre-Contrast	Real DCE-P3	84.13	60.96	0.208 ± .092	49.23 ± 16.15
CC-Net _{Any} P3	Real DCE-P3	61.17	37.80	0.190 ± .074	45.75 ± 13.74
Real Pre-Contrast	Real DCE-P4	100.27	77.31	0.199 ± .078	52.48 ± 12.96
CC-Net _{Any} P4	Real DCE-P4	47.13	60.80	0.198 ± .075	50.36 ± 14.26

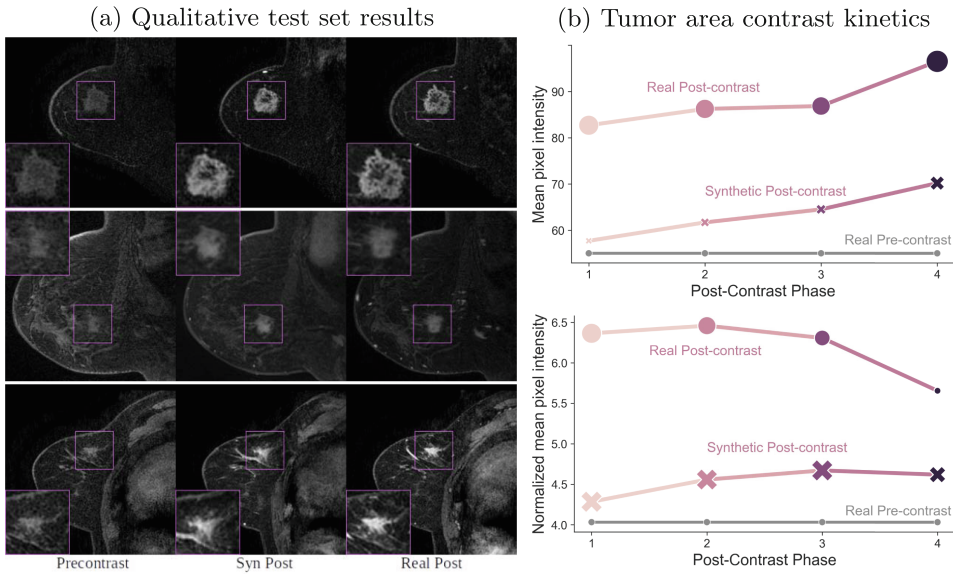


Fig. 3. Qualitative (a) and quantitative (b) test set contrast enhancement. In (b), marker size represents the standard deviation of the mean intensity within the tumor region averaged across all test cases. When normalized, tumor region mean intensity is divided by mean intensity of the remaining tumor-free pixels.

4 Discussion and Conclusion

We propose a multi-conditional latent diffusion model to translate pre-contrast into post-contrast images, thereby learning to highlight lesions by simulating their contrast uptake. We further condition the model on textual imaging meta-data and continuous time passed since pre-contrast acquisition and demonstrate its synthesis capabilities on multi-sequence breast DCE-MRI data. We further contribute the Fréchet radiomics distance (FRD), a novel radiology-specific image quality metric measuring the distance between real and synthetic distributions of extracted interpretable imaging biomarkers. We validate FRD demonstrating its correlation with image perturbation scales on both 3D and 2D data. In future investigations, we aim to generate images of multiple DCE-MRI time-points jointly and to map from the latent space of 3D autoencoders to the one from 2D-trained latent diffusion models. Further research avenues include confidence metrics to determine cases where synthetic contrast is sufficient, encoding additional clinical metadata using different biomedical text encoders, as well as exploring FRD feature compositions and variability. In conclusion, our work paves the way for practical applications of deep generative models in MRI as a screening modality for unsupervised tumor detection and localization from pre-contrast MRI. It further constitutes a step towards improved treatment of patient populations where invasive contrast agent injection is contraindicated.

Acknowledgements. This study has received funding from the European Union’s Horizon research and innovation programme under grant agreement No 952103 (EuCanImage) and No 101057699 (RadioVal). It was further partially supported by the project FUTURE-ES (PID2021-126724OB-I00) and by grant FJC2021-047659-I from the Ministry of Science and Innovation of Spain. RO acknowledges a research stay grant from the Helmholtz Information and Data Science Academy (HIDA).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Caballo, M., Sanderink, W.B., Han, L., Gao, Y., Athanasiou, A., Mann, R.M.: Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI. *Journal of Magnetic Resonance Imaging* **57**(1), 97–110 (2023)
2. Chambon, P., Bluethgen, C., Langlotz, C.P., Chaudhari, A.: Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133* (2022)
3. European Medicines Agency (EMA): EMA’s final opinion confirms restrictions on use of linear gadolinium agents in body scans. <https://www.ema.europa.eu/> (2023), online; accessed 06 August 2023
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680 (2014)

5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems* **30** (2017)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
7. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baekler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports* **13**(1), 7303 (2023)
8. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. *arXiv preprint [arXiv:2402.05210](https://arxiv.org/abs/2402.05210)* (2024)
9. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48**(4), 441–446 (2012)
10. Müller-Franzes, G., Huck, L., Tayebi Arasteh, S., Khader, F., Han, T., Schulz, V., Dethlefsen, E., Kathner, J.N., Nebelung, S., Nolte, T., et al.: Using Machine Learning to Reduce the Need for Contrast Agents in Breast MRI through Synthetic Images. *Radiology* **307**(3), e222211 (2023)
11. Olchowy, C., Cebulski, K., Łasecki, M., Chaber, R., Olchowy, A., Kałwak, K., Zaleska-Dorobisz, U.: The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-a systematic review. *PloS one* **12**(2), e0171704 (2017)
12. Osuala, R., Joshi, S., Tsirikoglou, A., Garrucho, L., Pinaya, W.H., Diaz, O., Lekadir, K.: Pre-to Post-Contrast Breast MRI Synthesis for Enhanced Tumour Segmentation. *arXiv preprint [arXiv:2311.10879](https://arxiv.org/abs/2311.10879)* (2023)
13. Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al.: medigan: a Python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging* **10**(6), 061403–061403 (2023)
14. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosi, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., et al.: Generative AI for Medical Imaging: extending the MONAI Framework. *arXiv preprint [arXiv:2307.15208](https://arxiv.org/abs/2307.15208)* (2023)
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10684–10695 (2022)
16. Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer* **119**(4), 508–516 (2018)
17. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*. pp. 2256–2265. PMLR (2015)
18. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: *International Conference on Learning Representations* (2021), <https://openreview.net/forum?id=St1giarCHLP>
19. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational

- radiomics system to decode the radiographic phenotype. *Cancer Research* **77**(21), e104–e107 (2017)
20. Woodland, M., Taie, M.A., Silva, J.A.M., Eltaher, M., Mohn, F., Shieh, A., Castelo, A., Kundu, S., Yung, J.P., Patel, A.B., et al.: Importance of Feature Extraction in the Calculation of Fréchet Distance for Medical Imaging. arXiv preprint [arXiv:2311.13717](https://arxiv.org/abs/2311.13717) (2023)
 21. Wu, S., Berg, W.A., Zuley, M.L., Kurland, B.F., Jankowitz, R.C., Nishikawa, R., Gur, D., Sumkin, J.H.: Breast mri contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer. *Breast Cancer Research* **18**, 1–10 (2016)
 22. Xing, X., Felder, F., Nan, Y., Papanastasiou, G., Walsh, S., Yang, G.: You don't have to be perfect to be amazing: Unveil the utility of synthetic images. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. pp. 13–22. Springer Nature Switzerland, Cham (2023)
 23. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3836–3847 (2023)
 24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
 25. Zhang, T., Han, L., D'Angelo, A., Wang, X., Gao, Y., Lu, C., Teuwen, J., Beets-Tan, R., Tan, T., Mann, R.: Synthesis of Contrast-Enhanced Breast MRI Using Multi-b-Value DWI-based Hierarchical Fusion Network with Attention Mechanism. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. pp. 79–88. Springer Nature Switzerland, Cham (2023)

4.1. Contribution: MICCAI 2024 (CCNet & FRD)

Chapter 5

Generative Models for Patient Privacy Preservation

In medical imaging, deep learning models have become invaluable for tasks such as cancer detection, diagnosis, and treatment planning. However, training these models requires large quantities of patient data, raising substantial concerns about data privacy. These concerns are especially pronounced as deep learning models, by design, can inadvertently memorize and leak sensitive patient information [203, 13, 86], making compliance with stringent data protection regulations more challenging. The General Data Protection Regulation (GDPR) [66] as one such regulation mandates robust safeguards for personal data, including the "right to be forgotten," which allows patients to request the deletion of their data, potentially causing extensive model retraining efforts. This chapter addresses these challenges by presenting, comparing, combining, and benchmarking privacy-preserving training of deep learning models with application to the clinical problem of breast cancer diagnosis based on detected lesions on mammograms.

5.1 Enhancing the Utility of Privacy-Preserving Cancer Classification using Synthetic Data

Summary

In this paper, we address the challenge of privacy-preserving deep learning combining and comparing different techniques with breast cancer classification performance as

5.1. Contribution: MICCAI DeepBreath Workshop 2024 (MCGAN)

downstream task. This study evaluates two main privacy-preserving approaches: Differentially Private Stochastic Gradient Descent (DP-SGD) [1] and synthetic data generation through a malignancy-conditioned generative adversarial network (MCGAN). The MCGAN produces realistic, labeled synthetic mammography images (benign or malignant), enabling model training without direct patient data.

To assess the effectiveness of these privacy-preserving methods, the study applies them to a Swin Transformer model [140] for breast cancer classification and benchmarks performance using two datasets, CBIS-DDSM [128] and BCDR [143], the latter being integrated into the study as external test set. Through various training setups, including real data alone, synthetic data augmentation, and DP-SGD at different privacy levels, it is shown that synthetic data can improve classification accuracy especially under stronger privacy protection. The findings suggest that combining DP-SGD with high-quality synthetic data offers an efficient and compliant solution for developing privacy-preserving diagnostic models for clinical practice.

Main Contributions:

- **Breast Cancer Classification Framework:** Developed a transformer-based model for breast cancer classification, achieving robust performance with privacy-preserving techniques, and validated on two benchmark breast cancer datasets, CBIS-DDSM [128] and BCDR [143].
- **Synthetic Data Generation:** Introduced and validated a malignancy-conditioned GAN (MCGAN) to generate synthetic benign and malignant breast masses, enabling privacy-preserving model training and supporting effective data quality for classification tasks.
- **Privacy-Utility Tradeoff Analysis:** Conducted a comprehensive privacy-utility analysis evaluating various differential privacy guarantees and finding that combining differential privacy (DP-SGD) with synthetic data improves performance, enabling privacy-preserving model fine-tuning without extensive retraining.

Contributions | The dissertation author led the conceptualization and development of the study including the manuscript preparation as well as the ideation, design, implementation, validation, and analysis of the malignancy-conditioned GAN, the differential privacy module, as well as the mass classification framework. Lang, D.M., Riess, A., Kaissis, G., Szafranowska, Z., Skorpuko, G., Diaz, O., Schnabel, J.A., Lekadir,

K. provided critical feedback and assisted in manuscript preparation. Szafranowska, Z., Skorpuko, G. helped in the implementation of the mass classification framework.

Declaration of Doctoral Use | None of the co-authors have used, either implicitly or explicitly, this publication for the development of their own doctoral thesis, nor do they intend to do so in the future.

Source Code | The reproducible source code of this work is made readily available at https://github.com/RichardObi/mammo_dp.

Publication | Conference: Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. First Deep Breast Workshop, Deep-Breath 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 10, 2024, Proceedings, Series ISSN: 0302-9743, Publisher: Springer Cham

Copyright Statement | Reproduced with permission from Springer Nature (License Number: 5985211068940). This contribution has been accepted for publication as: "Osuala, R., Lang, D.M., Riess, A., Kaissis, G., Szafranowska, Z., Skorpuko, G., Diaz, O., Schnabel, J.A., Lekadir, K. (2024). Enhancing the Utility of Privacy-Preserving Cancer Classification using Synthetic Data. *In Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (pp. 54-64)*. Cham: Springer Nature Switzerland." https://doi.org/10.1007/978-3-031-77789-9_6



Enhancing the Utility of Privacy-Preserving Cancer Classification Using Synthetic Data

Richard Osuala^{1,2,3} (✉), Daniel M. Lang^{2,3}, Anneliese Riess^{2,3},
Georgios Kaissis^{2,3,4}, Zuzanna Szafranowska¹, Grzegorz Skorupko¹,
Oliver Diaz^{1,5}, Julia A. Schnabel^{2,3,6}, and Karim Lekadir^{1,7}

¹ Departament de Matemàtiques i Informàtica, Universitat de Barcelona,
Barcelona, Spain

richard.osuala@ub.edu

² Helmholtz Center Munich, Munich, Germany

³ Technical University of Munich, Munich, Germany

⁴ Imperial College London, London, UK

⁵ Computer Vision Center, Bellaterra, Spain

⁶ Kings College London, London, UK

⁷ Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract. Deep learning holds immense promise for aiding radiologists in breast cancer detection. However, achieving optimal model performance is hampered by limitations in availability and sharing of data commonly associated to patient privacy concerns. Such concerns are further exacerbated, as traditional deep learning models can inadvertently leak sensitive training information. This work addresses these challenges exploring and quantifying the utility of privacy-preserving deep learning techniques, concretely, (i) differentially private stochastic gradient descent (DP-SGD) and (ii) fully synthetic training data generated by our proposed malignancy-conditioned generative adversarial network. We assess these methods via downstream malignancy classification of mammography masses using a transformer model. Our experimental results depict that synthetic data augmentation can improve privacy-utility tradeoffs in differentially private model training. Further, model pretraining on synthetic data achieves remarkable performance, which can be further increased with DP-SGD fine-tuning across all privacy guarantees. With this first in-depth exploration of privacy-preserving deep learning in breast imaging, we address current and emerging clinical privacy requirements and pave the way towards the adoption of private high-utility deep diagnostic models. Our reproducible codebase is publicly available at <https://github.com/RichardObi/mammo-dp>.

Keywords: Differential Privacy · Generative Models · Breast Imaging

1 Introduction

Breast cancer accounts for staggering estimates of 684.000 deaths and 2,26 million new cases worldwide per year [11]. Part of this burden could be reduced

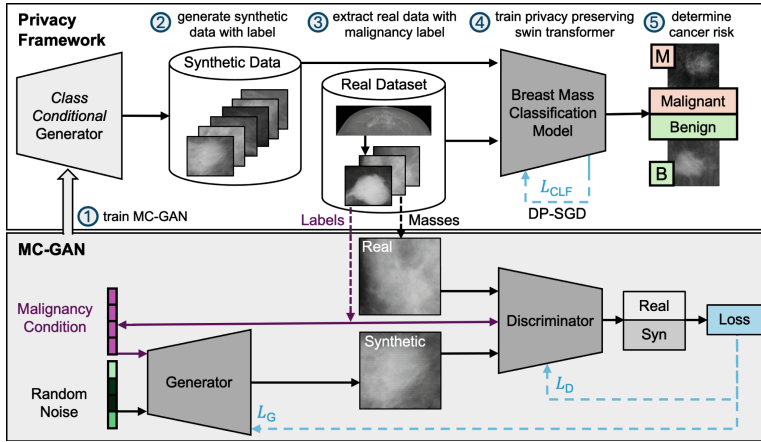


Fig. 1. Overview of our privacy-preserving deep learning pipeline and malignancy-conditioned generative adversarial network (MCGAN).

through earlier detection and timely treatment. Screening mammography is a cornerstone for early detection and further associated with a reduction in breast cancer mortality [21]. Recent literature emphasizes the potential of deep learning-based computer-aided diagnosis (CAD) [15, 22, 24, 30], e.g., demonstrating that a symbiosis of deep learning models with radiologist assessment yields the highest breast cancer detection performances [21]. However, training deep learning models on patient data poses a risk of leakage of sensitive person-specific information during and after training [24], as models have the capacity to memorise sufficient information to allow for high-fidelity image reconstruction [3, 13]. To avoid such leakage of private patient information, data needs to be protected during model training, in particular when the objective is to develop models to be used in clinical practice or shared among entities. Furthermore, international data protection regulations grant patients the right to request the removal of their information from data holders. For instance, point (b) of article 17(1) of the EU General Data Protection Regulation (GDPR) [9] stipulates that data subjects have a “right to be forgotten”. Given, for instance, the proven possibility of reconstructing training data given a model’s weights [3, 13], these rights can extend to the removal of patient-specific information from already trained deep learning models [29]. However, it is known to be difficult to “reliably” and “provably” remove patient information—present in only one or few specific training data points—from already trained model weights [29]. A generic and verifiable alternative is given by the removal of a patient’s data point from the training data and retraining of the respective model with the reminder of the dataset. This procedure is not only likely to have negative impacts on the performance of algorithms, but also emerges as a deterrence and risk for hospitals to adopt deep learning models, due to extensive economic, organisational, and environmental costs caused by retraining. Anticipating patient consent withdrawals, costly retraining can be avoided by demonstrating that deep learning model

weights do not include personally identifiable information (PII) about any specific patient. To this end, a powerful technique to ensure privacy during model training is given by Differentially Private Stochastic Gradient Descent (DP-SGD) [1], which quantifiably reduces the effect each single training sample can have on the resulting model weights. Furthermore, privacy-preservation can also be achieved by diagnostic models exclusively trained on synthetic data, which is not (unambiguously) attributable to any specific patient but rather contains anonymous samples representing the essence of the dataset [12, 24]. The caveat of both DP-SGD and synthetic data strategies is, however, that they generally lead to a reduction in model performance, known as the privacy-utility trade-off. Investigating this trade-off in the realm of breast imaging, our core contributions are summarised as follows:

- We design and validate a transformer model, achieving promising performance as a backbone for privacy-preserving breast mass malignancy classification.
- We propose and validate a conditional generative adversarial network capable of differentiating between benign and malignant breast mass generation.
- We empirically quantify privacy-utility-tradeoffs in mass malignancy classification, assessing various differential privacy guarantees, and further combine and compare them with training on synthetic data.

2 Methods and Materials

Datasets and Preprocessing

We use the open-access Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset [16], which consists of 891 scanned film mammography cases with segmented masses with biopsy-proven malignancy status. After extracting mass images from craniocaudal view (CC) and mediolateral oblique (MLO) views, we follow the predefined per-patient train-test split [16], allocating 1296 mass images for training and 402 (245 benign, 157 malignant) mass images to testing. We further divided this training set randomly per-patient into a training (1104 mass images, 525 malignant) and a validation set (192 mass images, 102 malignant). As external test set, we further adopt the publicly available BCDR cohort [19], which comprises 1010 patients, totalling 1493 lesions (639 masses) with contours and biopsy information from both digital mammograms (BCDR-DM) and film mammograms (BCDR-FM). Our final BCDR test set contains 1106 mass images extracted from CC and MLO views, 486 of which are malignant and 620 benign. To obtain mass patches from the mammograms, the lesion contour information is used to define bounding boxes, which enclose the mass. We then create a square patch around each bounding box with a minimum length and width of 128 pixels. Next, we increase the patch size by adding a margin of 60 pixels in each direction, before extracting the resulting patch, and resizing it to a pixel dimensions of 128×128 using inter-area interpolation. For classification, the mass patches are further resized to 224×224 px maintaining image ratios, and stacked to 3 channels. Models were

trained on either a single 8 GB NVIDIA RTX 2080 Super or 48 GB RTX A6000 GPU using PyTorch and opacus [31] for DP-SGD.

Cancer Classification Transformer Model

Given its reported high performance on classifying the presence of a lesion in mammography patches [30] and its shifted window mechanism, allowing to effectively attend to shapes of varying sizes, we adopt a swin transformer (Swin-T) [17] as cancer classification model, to distinguish between benign and malignant masses. We initialize ImageNet-pretrained [6] network weights and, after following the Swin-T hyperparameter setup [17] (stride, window size), we adjust the last fully-connected layer of the swin transformer reinitializing it with two output nodes each one outputting the logits for one of our respective classes (i.e., malignant or benign). We only set the parameters of the adjusted fully-connected layer as trainable and apply a learning rate of $1e-5$. A weight decay of $1e-8$ is used following the fine-tuning experiment described in [17]. Furthermore, an adamw optimizer [20], label smoothing of 0.1, and a batch size of 128 are used. During training, random horizontal and vertical flips are applied as data augmentation and a cross entropy loss is backpropagated. Training for 300 epochs using a cross entropy loss function, the model from the epoch with the lowest area under the precision-recall curve (AUPRC) on the validation set is selected for testing.

Malignancy-Conditioned Generative Adversarial Network

Going beyond unconditional mass synthesis in the literature [2, 30], we propose a malignancy conditioned generative adversarial network (MCGAN) to control the generation of either benign or malignant synthetic breast masses. In general, GANs consist of a generator (G) and a discriminator (D) network, which engage in a two-player zero-sum game, where G generates synthetic samples that D strives to distinguish from real ones [12]. We design G and D as deep convolutional neural networks [27] and, as shown in Fig. 1, integrate class-conditional information [23]. To this end, we extract the histopathology report’s biopsy information for each mass from the metadata, and convert it into a discrete malignancy label. Then, we transform this label into a multi-dimensional embedding vector to either represent the (a) malignant or (b) benign class, before passing it through a fully-connected layer yielding a representation with the corresponding dimensionality to concatenate it to the generator input (100 dim noise vector) and to the discriminator input (128×128 input image). As D learns to associate class labels with patterns in the input images, it has to learn whether or not a given class corresponds to a given synthetic sample. Furthermore, as the discriminator loss is backpropagated into the generator, G is forced to synthesize samples corresponding to the provided class condition. This results in G learning a conditional distribution based on the value function

$$\min_G \max_D V(D, G) = \min_G \max_D [\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x|y)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z|y)))]].$$

Optimizing the discriminator via binary cross-entropy [12], we define its loss in a class-conditional setup as

$$L_{D_{\text{MCGAN}}} = -\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x|y)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z|y)))].$$

We train our MCGAN on the CBIS-DDSM training data, applying random horizontal ($p = 0.5$) and vertical ($p = 0.5$) flipping as well as random cropping with resizing, where the resize scale ranges from 0.9 to 1.1 and aspect ratio from 0.95 to 1.1. We further include one-sided label smoothing [27] in a range of [0.7, 1.2]. Following [2], we employ a discriminator convolutional kernel size of 6 and a generator kernel size of 4. We observe that this reduces checkerboard artefacts as D’s field-of-view now requires G to create realistic transitions between the kernel-sized patches in the image. MCGAN is trained for 10k epochs with a batch size of 16. Based on the best quality-diversity tradeoff, we select the model from epoch 1.4k after qualitative visual assessment of generated samples.

Patient Privacy Preservation Framework

Privacy protection is an ethical norm and legal obligation, e.g. granting patients the right of their (retrospective) removal from databases [9]. Since (biomedical) deep learning models are vulnerable to information leakage, e.g. sensitive patient attributes [3, 13, 29], they can be affected by such (and future) regulations. However, privacy-preserving techniques can be integrated into deep learning frameworks and, to some extent, avoid compromising confidential data. For instance, (i) model training with DP-SGD [1] or (ii) training exclusively on synthetic data.

From a legal perspective, models trained on only synthetic data remain unaffected by patient consent withdrawal if “relatedness” between the data and the data subject cannot be established, or if “personal data has been rendered synthetic in such a manner that the data subject is no longer identifiable” [18] e.g., according to article 4(1) and recital 26 of the GDPR [9]. It is to be noted that in the “acceptable-risk” legal interpretation, a data subject’s re-identification risk is reduced to an “acceptable” level rather than fully eradicated [18]. Hence, this interpretation enables approaches such as synthetic data and/or Differential Privacy (DP) model training to be used as legally compliant privacy preservation methods despite not guaranteeing a “zero-risk” of patient re-identification.

DP is a mathematical framework that allows practitioners to provide (worst-case scenario) theoretical privacy guarantees for an individual sharing their data to train a deep learning model. Consider two databases (e.g., containing image-label pairs), we call them adjacent if they differ in a single data point, i.e., one image is present in one database but not in the other. Then, a randomised mechanism $\mathcal{M}: \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is said to satisfy (ϵ, δ) -differential privacy, if for any two adjacent databases $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$, $\Pr[\mathcal{M}(d) \in S] \leq e^\epsilon \Pr[\mathcal{M}(d') \in S] + \delta$ holds. ϵ and δ bound a single data point’s influence on a model’s output (e.g. the models’ weights or predictions). Thus, the smaller the value of these parameters, the higher the model’s privacy and the harder it is for an attacker to retrieve information

about any training data point. DP-SGD [1] is the DP variant of the well-known SGD algorithm, and facilitates the training of a model under DP conditions. In particular, a model trained under (ϵ, δ) -DP is robust to post-processing, meaning only using its output for further computations also satisfies (ϵ, δ) -DP. Moreover, the choice of these parameters is application-dependent and normative [5] and varies strongly across real-world deployments [7]. In the case of mammography, multiple lesions of the same patient are available in the datasets, i.e. one from the CC view and one from the MLO view. Therefore, to preserve the privacy of one patient it is necessary to protect all their data points (i.e. all images). In such a case, DP group privacy is used to estimate a patient’s DP privacy guarantee. However, for simplicity, in our subsequent experiments, we provide image-level privacy guarantees rather than per patient.

3 Experiments and Results

Experimental Setup		Metrics		
Dataset 1	Dataset 2	FID _{Img} ↓	FID _{Rad} ↓	FRD ↓
SynMCGAN	RealDDSM	58.00±0.72	0.81±.013	18.12±1.01
RealDDSM	RealDDSM	29.25±0.82	0.31±.019	3.48±.352
SynMCGAN	SynMCGAN	20.90±0.16	0.32±.012	0.57±.094
RealDDSM	RealBCDR	156.43±14.3	3.88±.351	277.63±39.0

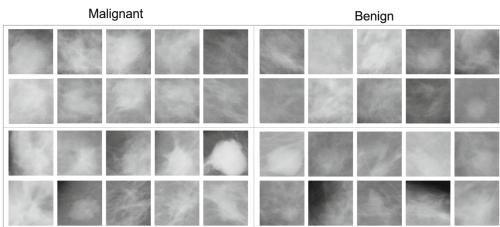


Fig. 2. Qualitative and quantitative synthesis results: Images are randomly selected malignant and benign real (CBIS-DDSM [16]) and MCGAN-generated masses. ImageNet [6] and RadImageNet [22, 26] based FID [14] and FRD [25] scores are reported as mean $\pm$ standard deviation based on 3 subsets randomly sampled per patient ($N_{\text{real}} \approx 360$, $N_{\text{syn}} \approx 3240$). Row 4 indicates an BCDR-based [19] upper bound for comparison with synthetic data metrics in row 1.

Synthetic Data Evaluation. Qualitatively assessing the synthetic images in Fig. 2, it is not readily possible to distinguish synthetic from real masses in terms of image fidelity or diversity. We note the absence of clear visual indicators to distinguish between malignant and benign images for both real and synthetic images. This is in line with the difficulty of determining the malignancy of a mammographic lesion shown by high clinical error rates and inter-observer variability [8]. However, results for training our malignancy classification model on only synthetic data (see *Syn* and *SynPre* in Table 1) show that the synthetic data captures the conditional distribution effectively generating either malignant or benign masses. Both, vanilla ImageNet-based Fréchet Inception Distance (FID) [6, 14] and radiology domain-specific RadImageNet-based FID [22, 26], concur that the synthetic data ($\text{FID}_{\text{Img}} = 58 \pm .72$) is substantially closer to the real CBIS-DDSM [16] distribution compared to BCDR [19] ($\text{FID}_{\text{Img}} = 156.43 \pm 1.43$). This is even more pronounced when comparing the

Table 1. Results for within-domain (CBIS-DDSM [16]) and out-of-domain (BCDR [19]) breast cancer malignancy classification masses extracted from mammograms. *Syn* indicates 3k synthetic images being part of the fine-tuning training data, while *Syn-Pre* represents pretraining all trainable model params with those 3k synthetic images (without DP guarantee), before fine-tuning the last two layers on real data with DP guarantee (*RealFT*). AUROC and AUPRC are reported as mean $\pm$ std based on 3 random seed runs. Best results in bold.

Experimental Setup			CBIS-DDSM [16]		BCDR [19]	
Model	ε	δ	AUROC $\uparrow$	AUPRC $\uparrow$	AUROC $\uparrow$	AUPRC $\uparrow$
SwinT _{Real}	∞	∞	0.778 $\pm$.001	0.85 $\pm$.001	0.695 $\pm$.002	0.726 $\pm$.003
SwinT _{Syn}	∞	∞	0.597 $\pm$.011	0.696 $\pm$.011	0.566 $\pm$.064	0.602 $\pm$.048
SwinT _{SynPre}	∞	∞	0.639 $\pm$.016	0.733 $\pm$.001	0.622 $\pm$.032	0.660 $\pm$.017
SwinT _{Real}	1	$1e^{-4}$	0.525 $\pm$.043	0.640 $\pm$.030	0.487 $\pm$.020	0.549 $\pm$.020
SwinT _{Real+Syn}	1	$1e^{-4}$	0.553$\pm$.040	0.665$\pm$.025	0.521$\pm$.023	0.573$\pm$.024
SwinT _{SynPre+RealFT}	$\infty-1$	$\infty-1e^{-4}$	0.661 $\pm$.018	0.741 $\pm$.007	0.637 $\pm$.026	0.67 $\pm$.0013
SwinT _{Real}	6	$1e^{-4}$	0.572 $\pm$.031	0.679 $\pm$.019	0.532 $\pm$.031	0.579 $\pm$.029
SwinT _{Real+Syn}	6	$1e^{-4}$	0.617$\pm$.013	0.708$\pm$.015	0.609$\pm$.027	0.647$\pm$.024
SwinT _{SynPre+RealFT}	$\infty-6$	$\infty-1e^{-4}$	0.677 $\pm$.014	0.752 $\pm$.009	0.647 $\pm$.022	0.679 $\pm$.009
SwinT _{Real}	12	$1e^{-4}$	0.596 $\pm$.023	0.702 $\pm$.013	0.559 $\pm$.033	0.600 $\pm$.030
SwinT _{Real+Syn}	12	$1e^{-4}$	0.624$\pm$.010	0.704$\pm$.012	0.625$\pm$.020	0.663$\pm$.012
SwinT _{SynPre+RealFT}	$\infty-12$	$\infty-1e^{-4}$	0.688 $\pm$.012	0.758 $\pm$.011	0.654 $\pm$.019	0.685 $\pm$.007
SwinT _{Real}	20	$1e^{-4}$	0.611 $\pm$.018	0.715$\pm$.012	0.581 $\pm$.028	0.618 $\pm$.026
SwinT _{Real+Syn}	20	$1e^{-4}$	0.630$\pm$.003	0.699 $\pm$.008	0.641$\pm$.018	0.685$\pm$.012
SwinT _{SynPre+RealFT}	$\infty-20$	$\infty-1e^{-4}$	0.697 $\pm$.012	0.763 $\pm$.012	0.659 $\pm$.017	0.689 $\pm$.006
SwinT _{Real}	60	$1e^{-4}$	0.622 $\pm$.014	0.721$\pm$.110	0.605 $\pm$.019	0.640 $\pm$.017
SwinT _{Real+Syn}	60	$1e^{-4}$	0.629$\pm$.002	0.694 $\pm$.005	0.650$\pm$.013	0.696$\pm$.007
SwinT _{SynPre+RealFT}	$\infty-60$	$\infty-1e^{-4}$	0.712 $\pm$.013	0.776 $\pm$.013	0.671 $\pm$.014	0.697 $\pm$.004

variation of extracted radiomics features for CBIS-DDSM to synthetic (FRD = 18.12) and BCDR (FRD = 277.63) images using the Fréchet Radiomics Distance (FRD) [25]. While this indicates desirable synthetic data fidelity, we also observe good diversity. The latter is shown by comparing subsets of the same datasets with each other, where the variation within the synthetic data (e.g., $\text{FID}_{\text{Rad}} = 0.32 \pm .12$) closely resembles the variation within the real CBIS-DDSM dataset (e.g., $\text{FID}_{\text{Rad}} = 0.31 \pm .19$). Notwithstanding less variation in radiomics imaging biomarkers within the synthetic data ($\text{FRD}_{\text{Syn}} = 0.57$ vs. $\text{FRD}_{\text{Real}} = 3.48$), this overall points to a valid coverage of the distribution and an absence of mode collapse.

Mass Malignancy Classification. As shown in Table 1, we conduct experiments with and without formal privacy guarantees. For scenarios where a formal privacy guarantee is not strictly required and, thus, synthetic data suffices as privacy mechanism, we compare the results of training SwinT on synthetic data (*Syn*) and on real data (*Real*) with DP-SGD. *Kaissis et al.* [15] defined

$\varepsilon = 6$ as suitable privacy budget for their medical imaging dataset. Compared to DP-SGD with $\varepsilon = 6$, synthetic data achieves better AUPRCs for within-domain tests on CBIS-DDSM ($\text{SwinT}_{\text{Syn}} = 0.696$ vs $\text{SwinT}_{\text{Real}(\varepsilon=6)} = 0.679$) and is on par for out-of-domain (ood) tests on BCDR ($\text{SwinT}_{\text{Syn}} = 0.602$ vs $\text{SwinT}_{\text{Real}(\varepsilon=6)} = 0.600$). However, training all SwinT layers using synthetic data (*SynPre*), achieves substantially better performance only approximated by DP results for $\varepsilon = 60$ for within-domain ($\text{SwinT}_{\text{SynPre}} = 0.733$ vs $\text{SwinT}_{\text{Real}(\varepsilon=60)} = 0.721$) and ood ($\text{SwinT}_{\text{SynPre}} = 0.66$ vs $\text{SwinT}_{\text{Real}(\varepsilon=60)} = 0.64$) tests. Further fine-tuning $\text{SwinT}_{\text{SynPre}}$ on real data using DP-SGD results in additional improvement across all privacy parameters for within-domain and ood testing. For instance, training $\text{SwinT}_{\text{SynPre+RealFT}}$ with $\varepsilon = 1$ results in an AUPRC of 0.74 and 0.67 for CBIS-DDSM and BCDR, respectively. To assess scenarios where a formal guarantee is required, we further compare DP-SGD training of SwinT on real data (*Real*) with DP-SGD training on a mix of real and synthetic data (*Real+Syn*). To this end, our experiments show that such synthetic data augmentation can improve the privacy-utility tradeoff. This is exemplified by $\text{SwinT}_{\text{Real+Syn}(\varepsilon=6)}$ accomplishing an AUPRC of 0.708 within-domain and 0.647 ood, while $\text{SwinT}_{\text{Real}(\varepsilon=6)}$ achieved 0.679 and 0.579, respectively. We further observe the trend that stricter privacy budgets (i.e., smaller ε) can be associated with more added performance of synthetic data as additional classification model training data.

4 Discussion and Conclusion

We introduce a privacy preservation framework based on differential privacy (DP) and synthetic data and apply it to the diagnostic task of classifying the malignancy of breast masses extracted from screening mammograms. We further propose, train, and evaluate a malignancy-conditioned generative adversarial network to generate a dataset of benign and malignant synthetic breast masses. Next, we train a swin transformer model on mass malignancy classification and assess, compare and combine training under DP and training on synthetic data. This analysis revealed that when training with DP, synthetic data augmentation can notably improve classification performance for within-domain and out-of-domain test cases. Apart from that, we show, across privacy mechanisms and across domains, that the performance of models pretrained on synthetic data can be further improved by DP fine-tuning on real data.

This finding is particularly important considering that synthetic data, if not directly attributable to any specific patient, can become a valid, legally compliant alternative to strict DP guarantees in clinical practice. Consequently, it is to be further investigated where and when deterministic mechanisms without formal DP guarantees can suffice to shield against different privacy attacks [4]. In particular, we motivate future work to analyse the extent to which the inherent properties of synthetic data generation algorithms can provide empirical protection against attacks. In this regard, a comparison of generation algorithms such as GANs [12] and denoising probabilistic diffusion models (DDPMs)

[28] can provide insights towards further improving conditional mass synthesis, while also enabling to quantify and compare the extent and effect of training data memorization in these models. A methodological alternative to our approach is to assess privacy-utility tradeoffs when training the generative model itself using DP-SGD [10, 24], resulting in formal privacy guarantees of the generated synthetic datasets. Thus, a further avenue to explore then lies within the question whether randomness inherent in randomised data synthesis algorithms (e.g., based on the noise in DDPMs or GANs) can be used to amplify the privacy of the DP versions of such synthesis algorithms, thereby potentially further enhancing privacy-utility tradeoffs. To this end, our study constitutes a crucial first step leading towards the clinical adoption of diagnostic deep learning models, enabling practical privacy-utility tradeoffs all while anticipating respective legal obligations and clinical requirements.

Acknowledgments. research and innovation programme under grant agreement No 952103 (EuCanImage) and No 101057699 (RadioVal). It was further partially supported by the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain. RO acknowledges a research stay grant from the Helmholtz Information and Data Science Academy (HIDA). GK received support from the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts under the Munich Centre for Machine Learning (MCML), from the German Ministry of Education and Research and the Medical Informatics Initiative as part of the PrivateAIM Project, from the Bavarian Collaborative Research Project PRIPREKI of the Free State of Bavaria Funding Programme “Artificial Intelligence – Data Science”, and from the German Academic Exchange Service (DAAD) under the Konrad Zuse School of Excellence for Reliable AI (RelAI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
2. Alyafi, B., Diaz, O., Marti, R.: DCGANs for realistic breast mass augmentation in x-ray mammography. In: Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314, p. 1131420. International Society for Optics and Photonics (2020)
3. Balle, B., Cherubin, G., Hayes, J.: Reconstructing training data with informed adversaries. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 1138–1156. IEEE (2022)
4. Cohen, A., Nissim, K.: Towards formalizing the gdpr’s notion of singling out. Proc. Natl. Acad. Sci. **117**(15), 8344–8352 (2020)
5. De, S., Berrada, L., Hayes, J., Smith, S.L., Balle, B.: Unlocking high-accuracy differentially private image classification through scale. arXiv preprint [arXiv:2204.13650](https://arxiv.org/abs/2204.13650) (2022)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)

7. Dwork, C., Kohli, N., Mulligan, D.: Differential privacy in practice: expose your epsilons! *J. Priv. Confidentiality* **9**(2) (2019)
8. Ekpo, E.U., Alakhras, M., Brennan, P.: Errors in mammography cannot be solved through technology alone. *Asian Pacific J. Can. Preven. APJCP* **19**(2), 291 (2018)
9. European Parliament and Council of European Union: General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of The European Parliament and of the Council. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679/> (2018)
10. Ghalebikesabi, S., et al.: Differentially private diffusion models generate useful synthetic images. arXiv preprint [arXiv:2302.13861](https://arxiv.org/abs/2302.13861) (2023)
11. Global Cancer Observatory: The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research. <https://gco.iarc.fr/> (2023), Accessed 17 Jan 2023
12. Goodfellow, I., et al.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014)
13. Haim, N., Vardi, G., Yehudai, G., Shamir, O., Irani, M.: Reconstructing training data from trained neural networks. arXiv preprint [arXiv:2206.07758](https://arxiv.org/abs/2206.07758) (2022)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint [arXiv:1706.08500](https://arxiv.org/abs/1706.08500) (2017)
15. Kaissis, G., et al.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**(6), 473–484 (2021)
16. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**(1), 1–9 (2017)
17. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021)
18. López, C.A.F.: On the legal nature of synthetic data. In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research* (2022)
19. Lopez, M.G., et al.: BCDR: a breast cancer digital repository. In: *15th International Conference on Experimental Mechanics*, vol. 1215 (2012)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *International Conference on Learning Representations* (2019)
21. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. *Nature* **577**(7788), 89–94 (2020)
22. Mei, X., et al.: RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* e210315 (2022)
23. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784) (2014)
24. Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., Lekadir, K.: Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging. *Med. Image Anal.* **84**, 102704 (2023)
25. Osuala, R., et al.: Towards learning contrast kinetics with multi-condition latent diffusion models. arXiv preprint [arXiv:2403.13890](https://arxiv.org/abs/2403.13890) (2024)
26. Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al.: medigan: a python library of pretrained generative models for medical image synthesis. *J. Med. Imaging* **10**(6), 061403 (2023)

27. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434) (2015)
28. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
29. Su, R., Liu, X., Tsaftaris, S.A.: Why patient data cannot be easily forgotten? In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part VIII, pp. 632–641. Springer (2022)
30. Szafranowska, Z., Osuala, R., Breier, B., Kushibar, K., Lekadir, K., Diaz, O.: Sharing generative models instead of private data: a simulation study on mammography patch classification. In: 16th International Workshop on Breast Imaging (IWBI2022), vol. 12286, pp. 169–177. SPIE (2022)
31. Yousefpour, A., et al.: Opacus: user-friendly differential privacy library in pytorch. arXiv preprint [arXiv:2109.12298](https://arxiv.org/abs/2109.12298) (2021)

Chapter 6

Sharing Generative Models to Overcome Data Scarcity Issues in Medical Imaging

Deep learning models often require large and diverse datasets to achieve accurate performance and desirable generalization capabilities. In contrast, sharing patient data across institutions is heavily restricted due to privacy regulations, limiting the development of models that perform well across varied clinical settings. An emerging alternative solution to this is the sharing of generative models, such as generative adversarial networks [83], which can generate synthetic data that mirrors the statistical properties of real patient data, thereby addressing both patient privacy while and data scarcity in medical imaging. For instance, in the dissertation author's previous work "Sharing Generative Models Instead of Private Data: A Simulation Study on Mammography Patch Classification" [206], it is demonstrated that GAN-generated mammography data from clinical center A can be effectively used to augment lesion classification training data from a second clinical center B resulting in classification performance increases in B . This effect was in particular visible in data-limited environments, and was present for both convolutional neural network [72, 127, 16] and Transformer-based models [215, 140], suggesting that generative model sharing not only enhances performance but also offers a scalable solution for deep learning model training in medical imaging potentially supporting more equitable access to high-quality healthcare AI.

The aforementioned findings, thus, motivate the introduction of a general framework

6.1. Contribution: JMI 2023 (medigan)

for efficiently sharing such generative models across centers and third parties. To this end, we propose the medigan library. medigan is designed as Python package and functions as a common platform for sharing and reusing of pretrained generative models, thus lowering the technical and computational barriers associated with creating synthetic medical data. By enabling researchers to access a broad selection of models trained on diverse datasets across modalities to address the challenges of data scarcity, privacy, and consistency in model training. The library allows to establish an operating standard for generative models and, thus, contributing to the progress towards robust, privacy-compliant deep learning solutions in healthcare and beyond.

6.1 medigan: a Python library of pretrained generative models for medical image synthesis

Summary:

The paper introduces medigan, a modular open-source Python library hosting multiple generative models across modalities and medical imaging domains that can produce synthetic, yet realistic, medical images that augment limited datasets. As training such models requires notable effort, hardware, and time, medigan bridges this gap by provides an accessible platform where researchers can search, compare, select, contribute, and utilize a variety of pretrained generative models. The library currently supports 20+ pretrained models across multiple modalities such as mammography, MRI, and endoscopy, enabling researchers to quickly generate diverse synthetic datasets without needing to train complex models from scratch. The paper demonstrates medigan’s utility through several applications: supporting privacy-preserving synthetic data sharing, enhancing clinical downstream task performance, and exploring standardized evaluation metrics like Fréchet Inception Distance (FID) for synthetic data quality assessment. While lowering technical barriers, medigan facilitates community-driven synthetic data sharing and paves the way towards standardized generative model usage in medical workflows.

Main Contributions:

- **Accessible Platform for Generative Models:** An comprehensive open-source Python library that allows for reuse, sharing, and contribution of pretrained generative models for synthetic medical image generation.
- **Comprehensive Model Repository:** Provides 20+ pretrained models spanning

multiple modalities (e.g., mammography, MRI, endoscopy), with metadata for each model to enable informed selection based on clinical and research needs.

- **Standardized Evaluation Tools:** Implements and tests synthetic data evaluation metrics, including domain-agnostic (ImageNet [50]) and domain-specific (RadImageNet [151]) Fréchet Inception Distance, and, based on extensive experiments, introduces best practices for reporting these metrics in medical imaging contexts.
- **Flexible Integration into AI Workflows:** Offers easy-to-use functions to integrate synthetic data generation directly into machine learning pipelines, supporting a variety of applications in medical image analysis and improving clinical task performance.
- **Demonstration of Performance Increase in Data-Limited Scenarios:** Showed that *medigan*'s synthetic data improves model performance for different clinical tasks ranging from mammography patch classification to tumor segmentation.

Contributions | The dissertation author led the conceptualization and development of the study including the ideation, design, development, implementation, experiments and publication of the *medigan* library. Apart from that, the dissertation author led the manuscript preparation, model integration, and the validation of the models in the *medigan* library using image quality metric definition and comparisons. Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Diaz, O., Lekadir, K. provided critical feedback and assisted in manuscript preparation. Skorupko, G. assisted in the implementation of the visualization features of the library and Lazrak, N. helped to design an initial proof-of-concept for model loading and inference. García, E., Joshi, S., Jouide, S., Kushibar, K., Garrucho, L., assisted in integrating some of the generative models into the library.

Declaration of Doctoral Use | None of the co-authors have used, either implicitly or explicitly, this publication for the development of their own doctoral thesis, nor do they intend to do so in the future.

Source Code | The reproducible source code of this work is made readily available at <https://github.com/RichardObi/medigan>. Up until the time of writing (22.04.2025), *medigan* had 154 stars and 19 forks on GitHub.

Python Package | The *medigan* library is available via the pip and conda package managers from the Python Packaging Index (PyPI) at <https://pypi.org/project/>

6.1. Contribution: JMI 2023 (medigan)

medigan/.

Publication | Journal: Journal of Medical Imaging, ISSN: 2329-4302, Publisher: SPIE, Journal Citation Reports (JCR) 2023, Impact Factor: 1.9

Copyright Statement | This work was originally published as: "Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., Diaz, O., Lekadir, K. (2023). medigan: a Python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging*, 10(6), 061403-061403."

<https://doi.org/10.1117/1.JMI.10.6.061403>

medigan: a Python library of pretrained generative models for medical image synthesis

Richard Osuala^{a,*}, Grzegorz Skorupko^a, Noussair Lazrak^a,
Lidia Garrucho^a, Eloy García^b, Smriti Joshi^a, Socayna Jouide^a,
Michael Rutherford^c, Fred Prior^c, Kaisar Kushibar^a,
Oliver Díaz^a and Karim Lekadir^a

^aUniversitat de Barcelona, Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM),
Facultat de Matemàtiques i Informàtica, Barcelona, Spain

^bUniversitat de Barcelona, Facultat de Matemàtiques i Informàtica, Barcelona, Spain

^cUniversity of Arkansas for Medical Sciences, Department of Biomedical Informatics,
Little Rock, Arkansas, United States

Abstract

Purpose: Deep learning has shown great promise as the backbone of clinical decision support systems. Synthetic data generated by generative models can enhance the performance and capabilities of data-hungry deep learning models. However, there is (1) limited availability of (synthetic) datasets and (2) generative models are complex to train, which hinders their adoption in research and clinical applications. To reduce this entry barrier, we explore generative model sharing to allow more researchers to access, generate, and benefit from synthetic data.

Approach: We propose *medigan*, a one-stop shop for pretrained generative models implemented as an open-source framework-agnostic Python library. After gathering end-user requirements, design decisions based on usability, technical feasibility, and scalability are formulated. Subsequently, we implement *medigan* based on modular components for generative model (i) execution, (ii) visualization, (iii) search & ranking, and (iv) contribution. We integrate pretrained models with applications across modalities such as mammography, endoscopy, x-ray, and MRI.

Results: The scalability and design of the library are demonstrated by its growing number of integrated and readily-usable pretrained generative models, which include 21 models utilizing nine different generative adversarial network architectures trained on 11 different datasets. We further analyze three *medigan* applications, which include (a) enabling community-wide sharing of restricted data, (b) investigating generative model evaluation metrics, and (c) improving clinical downstream tasks. In (b), we extract Fréchet inception distances (FID) demonstrating FID variability based on image normalization and radiology-specific feature extractors.

Conclusion: *medigan* allows researchers and developers to create, increase, and domain-adapt their training data in just a few lines of code. Capable of enriching and accelerating the development of clinical machine learning models, we show *medigan*'s viability as platform for generative model sharing. Our multimodel synthetic data experiments uncover standards for assessing and reporting metrics, such as FID, in image synthesis studies.

© The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JMI.10.6.061403](https://doi.org/10.1117/1.JMI.10.6.061403)]

Keywords: synthetic data; generative adversarial networks; Python; image synthesis; deep learning.

Paper 22262SSR received Oct. 5, 2022; accepted for publication Jan. 23, 2023; published online Feb. 20, 2023.

*Address all correspondence to Richard Osuala, Richard.Osuala@ub.edu

1 Introduction

1.1 Deep Learning and the Benefits of Synthetic Data

The use of deep learning has increased extensively in the last decade, thanks in part to advances in computing technology (e.g., data storage, graphics processing units) and the digitization of data. In medical imaging, deep learning algorithms have shown promising potential for clinical use due to their capability of extracting and learning meaningful patterns from imaging data and their high performance on clinically-relevant tasks. These include image-based disease diagnosis^{1,2} and detection,³ as well as medical image reconstruction,^{4,5} segmentation,⁶ and image-based treatment planning.^{7–9}

However, deep learning models need vast amounts of well-annotated data to reliably learn to perform clinical tasks, whereas, at the same time, the availability of public medical imaging datasets remains limited due to legal, ethical, and technical patient data sharing constraints.^{9,10} In the common scenario of limited imaging data, synthetic images, such as the ones illustrated in Fig. 1, are a useful tool to improve the learning of the artificial intelligence (AI) algorithm, e.g., by enlarging its training dataset.^{7,11,12} Furthermore, synthetic data can be used to minimize problems associated with domain shift, data scarcity, class imbalance, and data privacy.⁷ For instance, a dataset can be balanced by populating the less frequent classes with synthetic data during training (class imbalance). Further, as domain-adaptation technique, a dataset can be translated from one domain to another, e.g., from MRI to CT¹³ (domain shift). Regarding data privacy, synthetic data can be shared instead of real patient data to improve privacy preservation.^{7,14,15}

1.2 The Need of Reusable Synthetic Data Generators

Commonly, generative models are used to produce synthetic imaging data, with generative adversarial networks (GANs)¹⁶ being popular models of choice. However, the adversarial training scheme required by GANs and related networks is known to pose challenges in regard to (i) achieving training stability, (ii) avoiding mode collapse, and (iii) reaching convergence.^{17–19}

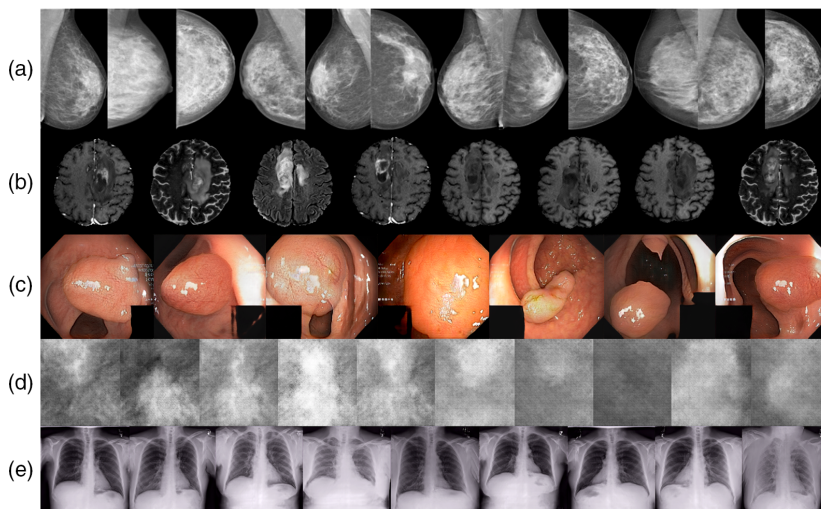


Fig. 1 Randomly sampled images generated by five *medigan* models ranging from (a) synthetic mammograms and (b) brain MRI to (c) endoscopy imaging of polyps, (d) mammogram mass patches, and (e) chest x-ray imaging. The models (a)–(e) correspond to the model IDs in Table 3, where (a) 3, (b) 7, (c) 10, (d) 12, and (e) 19.

Hence, the training process of GANs and generative models at large is nontrivial and requires a considerable time investment for each training iteration as well as specific hardware and a fair amount of knowledge and skills in the area of AI and generative modeling. Given these constraints, researchers and engineers often refrain from generating and integrating synthetic data into their AI training pipelines and experiments. This issue is further exacerbated by the prevailing need of training a new generative model for each new data distribution, which, in practice, often means that a new generative model has to be trained for each new application, use-case, and dataset.

1.3 Community-Driven Model Sharing and Reuse

We argue that a feasible solution to this problem is the community-wide sharing and reuse of pretrained generative models. Once successfully trained, such a model can be of value to multiple researchers and engineers with similar needs. For example, researchers can reuse the same model if they work on the same problem, conduct similar experiments, or evaluate their methods on the same dataset. We note that such reusing ideally is subject to previous inspection of generative model limitations with the model's output quality having qualified as suitable for the task at hand. The quality of a model's output data and annotations can commonly be measured via (a) expert assessment, (b) computation of image quality metrics, or (c) downstream task evaluation. In sum, the problem of synthetic data generation calls for a community-driven solution, where a generative model trained by one member of the community can be reused by other members of the community. Motivated by the absence of such a community-driven solution for synthetic medical data generation, we designed and developed *medigan* to bridge the gap between the need for synthetic data and complex generative model creation and training processes.

2 Background and Related Work

2.1 Generative Models

While discriminative models are able to distinguish between data instances of different kinds (label samples), generative models are able to generate new data instances (draw samples). In contrast to modeling decision boundaries in a data space, generative models model how data is distributed within that space. Deep generative models²⁰ are composed of multilayer neural networks to explicitly or implicitly estimate a probability density function (PDF) from a set of real data samples. After approximating the PDF from observed data points (i.e., learning the real data distribution), these models can then sample unobserved new data points from that distribution. In computer vision and medical imaging, synthetic images are generated by sampling such unobserved points from high-dimensional imaging data distributions. Popular deep generative models to create synthetic images in these fields include variational autoencoders,²¹ normalizing flows,^{22–24} diffusion models,^{25–27} and GANs.¹⁶ From these, the versatile GAN framework has seen the most widespread adoption in medical imaging to date.⁷ We, hence, center our attention on GANs in the remainder of this work but emphasize that contributions of other types of generative models are equally welcome in the *medigan* library.

2.2 Generative Adversarial Networks

The training of GANs comprises two neural networks, the generator network (G) and the discriminator network (D), as illustrated by Fig. 2 for the example of mammography region-of-interest patch generation. G and D compete against each other in a two-player zero-sum game defined by the value function shown in Eq. (1). Subsequent studies extended the adversarial learning scheme by proposing innovations of the loss function, G and D network architectures, and GAN applications by introducing conditions into the image generation process

$$\min_G \max_D V(D, G) = \min_G \max_D [\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))]. \quad (1)$$

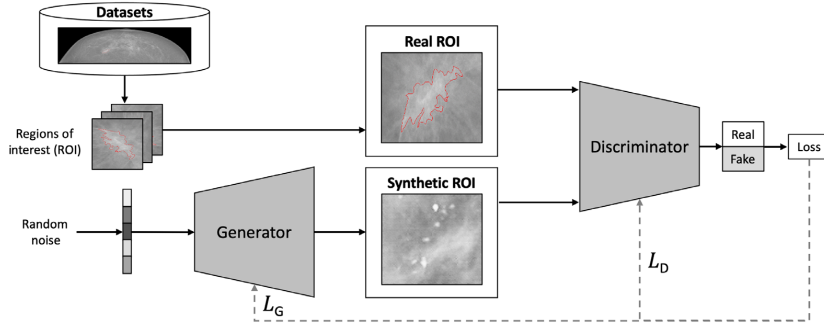


Fig. 2 The GAN framework. In this visual example, the generator network receives random noise vectors, which it learns to map to region-of-interest patches of full-field digital mammograms. During training, the adversarial loss is not only backpropagated to the discriminator as L_D but also to the generator as L_G . This particular architecture and loss function was used to train *medigan* models listed with IDs 1, 2, and 5 in Table 3.

2.2.1 GAN loss functions

Goodfellow et al.¹⁶ define the discriminator as a binary classifier classifying whether a sample x is either real or generated. The discriminator is, hence, trained via binary-cross entropy with the objective of minimizing the adversarial loss function shown in Eq. (2), which the generator, on the other hand, tries to maximize. In Wasserstein GAN (WGAN),²⁸ the adversarial loss function is replaced with a loss function based on the Wasserstein-1 distance between real and fake sample distributions estimated by D (alias “critic”). Gulrajani et al.²⁹ resolve the need to enforce a 1-Lipschitz constraint in WGAN via gradient penalty (WGAN-GP) instead of WGAN weight clipping. Equation (3) depicts the WGAN-GP discriminator loss with penalty coefficient λ and distribution $\mathbb{P}_{\tilde{x}}$ based on sampled pairs from (a) the real data distribution $\mathbb{P}_{\text{data}}$ and (b) the generated data distribution $\mathbb{P}_g$

$$L_{D_{\text{GAN}}} = -\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))], \quad (2)$$

$$L_{D_{\text{WGAN-GP}}} = E_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - E_{x \sim \mathbb{P}_{\text{data}}} [D(x)] + \lambda E_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} [(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2]. \quad (3)$$

In addition to changes to the adversarial loss, further studies integrate additional loss terms into the GAN framework. For instance, FastGAN³⁰ uses an additional reconstruction loss in the discriminator, which, for improved regularisation, is trained as self-supervised feature-encoder.

2.2.2 GAN network architectures and conditions

A plethora of different GAN network architectures has been proposed^{7,31} starting with a deep convolutional GAN (DCGAN)³² neural network architecture of both D and G . Later approaches, e.g., include a ResNet-based architecture as backbone²⁹ and progressively-grow the generator and discriminator networks during training to enable high-resolution image synthesis (PGGAN).³³

Another line of research has been focusing on conditioning the output of GANs based on discrete or continuous labels. For example, in cGAN this is achieved by feeding a label to both D and G ,³⁴ whereas in the auxiliary classifier GAN (AC-GAN), the discriminator additionally predicts the label that is provided to the generator.³⁵

Other models condition the generation process on input images^{36–40} unlocking image-to-image translation and domain-adaptation GAN applications. A key difference in image-to-image translation methodology is the presence (paired translation) or absence (unpaired translation) of corresponding image pairs in the target and source domain. Using an L1 reconstruction loss between target and source domain alongside the adversarial loss from Eq. (2), pix2pix³⁶ defines a common baseline model for paired image-to-image translation. For unpaired translation,

cycleGAN³⁷ is a popular approach, which also consists of an L1 reconstruction (cycle-consistency) loss between a source (target) image and a source (target) image translated to target (source) and back to source (target) via two consecutive generators.

A further methodological innovation includes SinGAN,⁴¹ which, based on only a single training image, learns to generate multiple synthetic images. This is accomplished via a multi-scale coarse-to-fine pipeline of generators, where a sample is passed sequentially through all generators, each of which also receives a random noise vector as input.

2.3 Generative Model Evaluation

One approach of evaluating generative models is by human expert assessment of their generated synthetic data. In medical imaging, such observer studies often enlist board-certified clinical experts such as radiologists or pathologists to examine the quality and/or realism of the synthetic medical images.^{42,43} However, this approach is manual, laborious and costly, and, hence, research attention has been devoted to automating generative model evaluation,^{44,45} including:

- i. Metrics for automated analysis of the synthetic data and its distribution, such as the inception score (IS)¹⁷ and Fréchet inception distance (FID).⁴⁶ Both metrics are popular in computer vision,³¹ whereas the latter also has seen widespread adoption in medical imaging.⁷

FID is based on a pretrained Inception⁴⁷ model (e.g., v1,⁴⁸ v3⁴⁷) to extract features from synthetic and real datasets, which are then fitted to multivariate Gaussians X (e.g., real) and Y (e.g., synthetic) with means μ_X and μ_Y and covariance matrices Σ_X and Σ_Y . Next, X and Y are compared via the Wasserstein-2 (Fréchet) distance (FD), as depicted as

$$FD(X, Y) = \|\mu_X - \mu_Y\|_2^2 + \text{tr}\left(\Sigma_X + \Sigma_Y - 2\left(\Sigma_X \Sigma_Y\right)^{\frac{1}{2}}\right). \quad (4)$$

- ii. Metrics that compare a synthetic image with a real reference image such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM).⁴⁹ Given the absence of corresponding reference images, such metrics are not readily applicable for unconditional noise-to-image generation models.
- iii. Metrics that compare the performance of a model on a surrogate downstream task with and without generative model intervention.^{7,14,50,51} For instance, training on additional synthetic data can increase a model's downstream task performance, thus, demonstrating the usefulness of the generative model that generated such data.

For the analysis of generative models in the present study, we discard (ii) due to its limitation of requiring specific reference images. We further deprioritize the IS from (i) due to its limited applicability to medical imagery stemming from it missing a comparison between real and synthetic data distributions combined with it having a strong bias on natural images via its ImageNet⁵²-pretrained Inception classifier as backbone feature extractor. Therefore, we focus on FID from (i) and downstream task performance (iii) as potential evaluation measures for medical image synthesis models in the remainder of this work.

2.4 Image Synthesis Tools and Libraries

Related libraries, such as pygan,⁵³ torchGAN,⁵⁴ vegans,⁵⁵ imaginaire,⁵⁶ TF-GAN,⁵⁷ PyTorch-GAN,⁵⁸ keras-GAN,⁵⁹ mimicry,⁶⁰ and studioGAN,³¹ have focused on facilitating the implementation, training, and comparative evaluation of GANs in computer vision (CV). Despite a strong focus on language models, the HuggingFace transformers library and model hub⁶¹ also contain a few pretrained computer vision GAN models. The GAN Lab⁶² provides an interactive visual experimentation tool to examine the training process and its data flows in GANs.

Specific to AI in medical imaging, Diaz et al.⁶³ provided a comprehensive survey of tools, libraries and platforms for privacy preservation, data curation, medical image storage, annotation, and repositories. Compared to CV, fewer GAN and AI libraries and tools exist in medical imaging. Furthermore, CV libraries are not always suited to address the unique challenges of

medical imaging data.^{63–65} For instance, pretrained generative models from computer vision cannot be readily adapted to produce medical imaging-specific outputs. The TorchIO library⁶⁴ addresses the gap between CV and medical image data processing requirements providing functions for efficient loading, augmentation, preprocessing, and patch-based sampling of medical imagery. The medical open network for AI (MONAI)⁶⁶ is a PyTorch-based⁶⁷ framework that facilitates the development of diagnostic AI models with tutorials for classification, segmentation, and AI model deployment. Further efforts in this realm include NiftyNet,⁶⁸ the deep learning tool kit (DLTK),⁶⁹ MedicalZooPytorch,⁷⁰ and nnDetection.⁷¹ The recent RadImageNet initiative⁷² shares baseline image classification models pretrained on a dataset designed as the radiology medical imaging equivalent to ImageNet.⁵²

To the best of our knowledge, no open-access software, tool, or library exists that targets reuse and sharing of pretrained generative models in medical imaging. To this end, we expect the contribution of our *medigan* library to be instrumental in enabling dissemination of generative models and increased adoption of synthetic data into AI training pipelines. As an open-access plug-and-play solution for generation of multipurpose synthetic data, *medigan* aims to benefit patients and clinicians by enhancing the performance and robustness of AI-based clinical decision support systems.

3 Method: The *medigan* Library

We contribute *medigan* as an open-source open-access MIT-licensed Python3 library distributed via the Python package index (Pypi) for synthetic medical dataset generation, e.g., via pretrained generative models. The metadata of *medigan* is summarized in Table 1. *medigan* accelerates research in medical imaging by flexibly providing (a) synthetic data augmentation and (b) pre-processing functionality, both readily integrable in machine learning training pipelines. It also allows contributors to add their generative models in a thought-through process and provides simplistic functions for end-users to search for, rank, and visualize models. The overview of *medigan* in Fig. 3 depicts the core functions demonstrating how end-users can (a) contribute a generative model, (b) find a suitable generative model inside the library, and (c) generate synthetic data with that model.

Table 1 Overview of *medigan* library information.

	Title	<i>medigan</i> metadata
1	Code version	v1.0.0
2	Code license	MIT
3	Code version control system	Git
4	Software languages	Python
5	Code repository	https://github.com/RichardObi/medigan .
6	Software package repository	Ref. 73
7	Developer documentation	Ref. 74
8	Tutorial	medigan quickstart (tutorial.ipynb)
9	Requirements for compilation	Python v3.6+
10	Operating system	OS independent. Tested on Linux, OSX, Windows.
11	Support email address	Richard.Osuala[at]gmail.com
12	Dependencies	tqdm, requests, torch, numpy, PyGithub, matplotlib (setup.py)

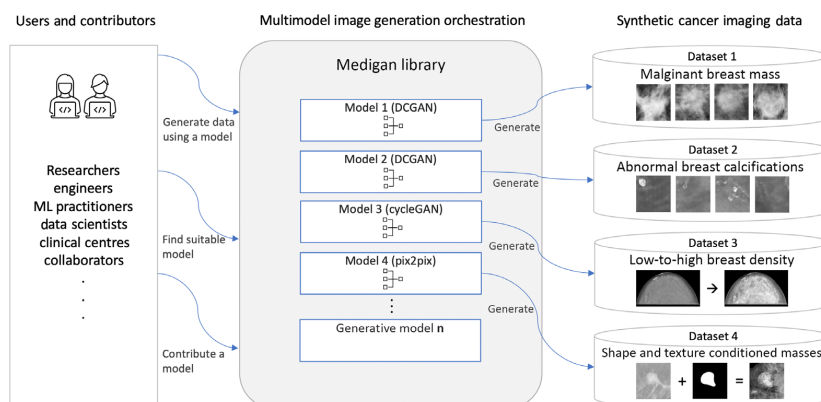


Fig. 3 Architectural overview of *medigan*. Users interact with the library by contributing, searching, and executing generative models, the latter shown here exemplified for mammography image generation with models with IDs 1 to 4 described in Table 3.

3.1 User Requirements and Design Decisions

End-user requirement gathering is recommended for the development of trustworthy AI solutions in medical imaging.⁷⁵ Therefore, we organized requirement gathering sessions with potential end-users, model contributors, and stakeholders from the EuCanImage Consortium, a large European H2020 project⁷⁶ building a cancer imaging platform for enhanced AI in oncology. Upon exploring the needs and preferences of medical imaging researchers and AI developers, respective requirements for the design of *medigan* were formulated to ensure usability and usefulness. For instance, the users articulated a clear preference for a user interface in the format of an importable package as opposed to a graphical user interface (GUI), web application, database system, or API. Table 2 summarizes key requirements and the corresponding design decisions.

3.2 Software Design and Architecture

medigan is built with a focus on simplicity and usability. The integration of pretrained models is designed as internal Python package import and offers simultaneously (a) high flexibility to and (b) low code dependency on these generative models. The latter allows the reuse of the same orchestration functions in *medigan* for all model packages.

Using object-oriented programming, the same `model_executor` class is used to implement, instantiate, and run all different types of generative model packages. To keep the library maintainable and lightweight, and to avoid limiting interdependencies between library code and generative model code, *medigan*'s models are hosted outside the library (on Zenodo) as independent Python modules. To avoid long initialization times upon library import, lazy loading is applied. A model is only loaded and its `model_executor` instance is only initialized if a user specifically requests synthetic data generation for that model. To achieve high cohesion,⁷⁹ i.e., keeping the library and its functions specific, manageable, and understandable, the library is structured into several modular components. These include the loosely-coupled `model_executor`, `model_selector`, and `model_contributor` modules.

The `generators` module is inspired by the facade design pattern⁸⁰ and acts as a single point of access to all of *medigan*'s functionalities. As single interface layer between users and library, it reduces interaction complexity and provides users with a clear set of readily extendable library functions. Also, the `generators` module increases internal code reusability and allows for combination of functions from other modules. For instance, a single function call can run the generation of samples by the model with the highest FID score of all models found in a keyword search.

Table 2 Overview of the key requirements gathered together with potential end-user alongside the respective design decisions taken toward fulfilling these requirements with *medigan*.

No	End-user requirement	Respective design decision
1	Instead of a GUI tool, <i>medigan</i> should be implemented as a platform-independent library importable into users' code.	Implementation of <i>medigan</i> as publicly accessible Python package distributed via PyPI.
2	It should support common frameworks for building generative models, e.g., PyTorch, ⁶⁷ TensorFlow, ⁷⁷ Keras. ⁷⁸	<i>medigan</i> is built framework-agnostic treating each model as separate Python package with freedom of choice of framework and dependencies.
3	The library should allow different types of generative models and generation processes.	<i>medigan</i> supports any type of data generation model including GANs, ¹⁶ VAEs, ²¹ flow-based, ²²⁻²⁴ diffusion, ²⁵⁻²⁷ and nondeep learning models.
4	The library should support different types of synthetic data.	<i>medigan</i> supports any type of synthetic data ranging from 2D and 3D images to image pairs, masks, and tabular data.
5	Sample generation functions should be easily integrable into diverse user code, pipelines, and workflows.	<i>medigan</i> 's generate function can (i) return samples, (ii) generate folders with samples, or (iii) return a model's generate function as callable.
6	User should be able to integrate <i>medigan</i> data in AI training via a data loader.	For each model, <i>medigan</i> supports returning a torch data loader readily integrable in AI training pipelines, combinable with other data loaders.
7	Despite using large deep learning models, the library should be as lightweight as possible.	Only the user-requested models are downloaded and locally imported. Thus, model dependencies are not part of <i>medigan</i> 's dependencies.
8	It should be possible to locally review and adjust a generative model of the library.	After download, a model's code and config are available for end-users to explore and adjust. <i>medigan</i> can also load models from local file systems.
9	The library should support both CPU and GPU usage depending on a user's hardware.	Contributed <i>medigan</i> models are reviewed and, if need be, enhanced to run on both GPU and CPU.
10	Version and source of the models that the library load should be transparent to the end-user.	Convention of storing <i>medigan</i> models on Zenodo, where each model's source code and version history is available.
11	There should be no need to update the version of the <i>medigan</i> package each time a new model is contributed.	<i>medigan</i> is designed independently of its model packages separately stored on Zenodo. Config updates do not require new <i>medigan</i> versions.

Table 2 (Continued).

No	End-user requirement	Respective design decision
12	Following, ⁷⁵ models are contributed in transparently and traceably, allowing quality and reproducibility checks.	Model contribution is traceable via version control. Adding models to <i>medigan</i> requires a config change via pull request.
13	The risk that the library downloads models that contain malicious code should be minimized.	Zenodo model uploads receive static DOIs. After verification, unsolicited uploads/changes do not affect <i>medigan</i> , which points to specific DOI.
14	License and authorship of generative model contributors should be clearly stated and acknowledged.	Separation of models and library allows freedom of choice of model license and transparent authorship reported for each model.
15	Each generative model in the library should be documented.	Each available model is listed and described in <i>medigan</i> 's documentation, in the readme, and also separately in its Zenodo entry.
16	The library should have minimal dependencies on the user side and should run on common end-user systems.	<i>medigan</i> has a minimal set of Python dependencies, is OS-independent, and avoids system and third-party dependencies.
17	Contributing models should be simple and at least partially automated.	<i>medigan</i> 's contribution workflow automates local model configuration, testing, packaging, Zenodo upload, and issue creation on GitHub.
18	If different models have the same dependency but with different versions, this should not cause a conflict.	Model dependency versions are specified in the config. <i>medigan</i> 's generate method can install unsatisfied dependencies, avoiding conflicts.
19	Any model in the library should be automatically tested and results reported to make sure all models work as designed.	On each commit to main, a CI pipeline automatically builds, formats, and lints <i>medigan</i> before testing all models and core functions.
20	The library should make the results of the models visible with minimal code required by end-users.	<i>medigan</i> 's simple visualization feature allows users to adjust a model's input latent vector for intuitive exploration of output diversity and fidelity.
21	The library should support large synthetic dataset generation on user machines with limited random-access memory.	For large synthetic dataset generation, <i>medigan</i> iteratively generates samples via small batches to avoid exceeding users' in-memory storage limits.
22	Users can specify model weights, model inputs, number, and storage location of the synthetic samples.	Diverging from defaults, users can specify (i) weights, (ii) number of samples (iii) return or store, (iv) store location, (v) optional inputs.

3.3 Model Metadata

The FID score and all other model information such as dependencies, modality, type, zenodo link, associated publications, and generate function parameters are stored in a single comprehensive model metadata json file. Alongside its searchability, readability, and flexibility, the choice of json as file format is motivated by its extendability to a nonrelational database. As a single source of model information, the *global.json* file consists of an array of model IDs, where under each model id the respective model metadata is stored. Toward ensuring model traceability as recommended by the FUTURE-AI consensus guidelines,⁷⁵ each model (on Zenodo) and its *global.json* metadata (on GitHub) are version-controlled with the latter being structured into the following objects.

- i. *execution*: contains the information needed to download, package, and run the model resources.
- ii. *selection*: contains model evaluation metrics and further information used to search, compare, and rank models.
- iii. *description*: contains general information and main details about the model such as title, training dataset, license, date, and related publications.

This *global.json* metadata file is retrieved, provided, and handled by the *config_manager* module once a user imports the *generators* module. This facilitates rapid access to a model's metadata given its *model_id* and allows one to add new models or model versions to *medigan* via pull request without requiring a new release of the library.

3.4 Model Search and Ranking

The number of models in *medigan* is expected to grow over time. Potentially this will lead to the foreseeable issue where users of *medigan* have a large number of models to choose from. Users likely will be uncertain which model best fits their needs depending on their data, modality, use-case, and research problem at hand and would have to go through each model's metadata to find the most suitable model in *medigan*. Hence, to facilitate model selection, the *model_selector* module implements model search and ranking functionalities. This search workflow is shown in Fig. 4 and triggered by running Code Snippet 1.

The *model_selector* module contains a search method that takes search operator (i.e OR, AND, or XOR) and a keyword search values list as parameters and recursively searches through the models' metadata. The latter is provided by the *config_manager* module.

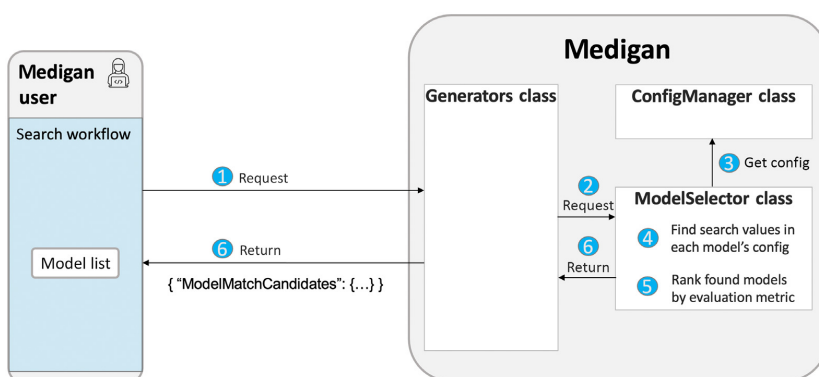


Fig. 4 The search workflow. A user sends a search query (1) to the generators class, which triggers a search (2) via the ModelSelector class. The latter retrieves the *global.json* model metadata/config dict (3), in which it searches for query values finding matching models (4). Next, the matched models are optionally also ranked based on a user-defined performance indicator (5) before being returned as list to the user.

Code Snippet 1: Searching for a model in *medigan*.

```

1. from medigan import Generators # import
2. generators = Generators() # init
3. values=['patches', 'mammography'] # keywords of search query
4. operator='AND' # all keywords are needed for match
5. results = generators.find_model(values, operator)

```

The `model_selector` populates a `modelMatchCandidates` object with `matchedEntry` instances each of which represents a potential model match to the search query. The `modelMatchCandidates` class evaluates which of it is associated model matches should be flagged as true match given the search values and search operator. The method `rank_models_by_performance` compares either all or specified models in *medigan* by a performance indicator such as FID. This indicator commonly is a metric that correlates with diversity, fidelity, or condition adherence to estimate the quality of generative models and/or the data they generate.⁷ The `model_selector` looks up the value for the specified performance indicator in the model metadata and returns a descendingly or ascendingly ranked list of models to the user.

3.5 Synthetic Data Generation

Synthetic data generation is *medigan*'s core functionality toward overcoming scarcity of (a) training data and (b) reusable generative model in medical imaging. Posing a low entry barrier for nonexpert users, *medigan*'s `generate` method is both simple and scalable. While a user can run it with only one line of code, it flexibly supports any type of generative model and synthetic data generation process, as illustrated in Table 3 and Fig. 1.

3.5.1 Generate workflow

An example of the usage of the `generate` method is shown in Code Snippet 2, which triggers the model execution workflow illustrated in Fig. 5. Further parameters of the `generate` method allow users to specify the number of samples to be generated (`num_samples`), if samples are returned as a list or stored on a disk (`save_images`), where they are stored (`output_path`), and whether model dependencies are automatically installed (`install_dependencies`). Optional model-specific inputs can be provided via the `**kwargs` parameter. These include for example, (i) a nondefault path to the model weights, (ii) a path to an input image folder for image-to-image translation models, (iii) a conditional input for class-conditional generative models, or (iv) the `input_latent_vector` as commonly used as model input in GANs.

Running the `generate` method triggers the `generators` module to initialize a `model_executor` instance for the user-specified generative model. The model is identified via its `model_id` as unique key in the `global.json` model metadata database, parsed and managed by the `config_manager` module. Using the latter, the `model_executor` checks if the required Python package dependencies are installed, retrieves the Zenodo URL and downloads, unzips, and imports the model package. It further retrieves the name of the internal data generation function inside the model's `__init__.py` script. As final step before calling this function, its parameters and their default values are retrieved from the metadata and combined with user-provided arguments. These user-provided arguments customize the generation process, which enables handling of multiple image generation scenarios. For instance, the aforementioned provision of the input image folder allows users to point to their own images to transform them using *medigan* models that are, e.g., pretrained for cross-modality translation. In the case of large dataset generation, the number of samples indicated by `num_samples` are chunked into smaller-sized batches and iteratively generated to avoid overloading the random-access memory available on the user's machine.

Table 3 Models currently available in *medigan*. Also, computed FID scores for each model in *medigan* are shown. The number of real samples used for FID calculation is indicated by #imgs. The *lower bound* FID_r is computed between a pair of randomly sampled sets of real data (real-real), whereas the *model* FID_s is computed between two randomly sampled sets of real and synthetic data (real-syn). The results for model 7 (Flair, T1, T1c, T2) and 21 (T1, T2) are averaged across modalities.

ID	Output	Modality	Model	Size	Training dataset	FID _{ImageNet} ^{47,52}			
						#imgs	Real-real	Real-syn	r _{FID}
1	Breast calcifications	Mammography	DCGAN	128 × 128	INbreast ⁸¹	1000	33.61	67.60	0.497
2	Breast masses	Mammography	DCGAN ¹¹	128 × 128	OPTIMAM ⁸²	1000	28.85	80.51	0.358
3	High/low density breasts	Mammography	CycleGAN ⁵⁰	1332 × 800	BCDR ⁸³	74	65.94	150.16	0.439
4	Breast masses with masks	Mammography	pix2pix	256 × 256	BCDR ⁸³	199	68.22	161.17	0.423
5	Breast masses	Mammography	DCGAN ¹⁴	128 × 128	BCDR ⁸³	199	68.22	180.04	0.379
6	Breast masses	Mammography	WGAN-GP ¹⁴	128 × 128	BCDR ⁸³	199	68.22	221.30	0.308
7	Brain tumors with masks	Cranial MRI	Inpaint GAN ⁸⁴	256 × 256	BRATS 2018 ⁸⁵	1000	30.73	140.02	0.219
8	Breast masses (mal/benign)	Mammography	C-DCGAN	128 × 128	CBIS-DDSM ⁸⁶	379	37.56	137.75	0.272
9	Polyps with masks	Endoscopy	PGGAN ⁵¹	256 × 256	HyperKvasir ⁸⁷	1000	43.31	225.85	0.192
10	Polyps with masks	Endoscopy	FastGAN ⁵¹	256 × 256	HyperKvasir ⁸⁷	1000	43.31	63.99	0.677
11	Polyps with masks	Endoscopy	SinGAN ⁵¹	≈250 × 250	HyperKvasir ⁸⁷	1000	43.31	171.15	0.253
12	Breast masses (mal/benign)	Mammography	C-DCGAN	128 × 128	BCDR ⁸³	199	68.22	205.29	0.332
13	High/low density breasts MLO	Mammography	CycleGAN ⁵⁰	1332 × 800	OPTIMAM ⁸²	358	65.75	101.09	0.650
14	High/low density breasts CC	Mammography	CycleGAN ⁵⁰	1332 × 800	OPTIMAM ⁸²	350	41.61	73.77	0.564
15	High/low density breasts MLO	Mammography	CycleGAN ⁵⁰	1332 × 800	CSAW ⁶⁸	192	74.96	162.67	0.461
16	High/low density breasts CC	Mammography	CycleGAN ⁵⁰	1332 × 800	CSAW ⁶⁸	202	42.68	98.38	0.434
17	Lung nodules	Chest x-ray	DCGAN	128 × 128	NODE21 ⁸⁹	1476	24.34	126.78	0.192
18	Lung nodules	Chest x-ray	WGAN-GP	128 × 128	NODE21 ⁸⁹	1476	24.34	211.47	0.115
19	Full chest radiograph	Chest x-ray	PGGAN	1024 × 1024	ChestX-ray14 ⁹⁰	1000	28.74	96.74	0.297
20	Full chest radiograph	Chest x-ray	PGGAN ⁹¹	1024 × 1024	ChestX-ray14 ⁹⁰	1000	28.33	52.17	0.543
21	Brain scans (T1/T2)	Cranial MRI	CycleGAN ⁸²	224 × 192	CrossModA 2021 ⁸³	1000	24.41	59.49	0.410

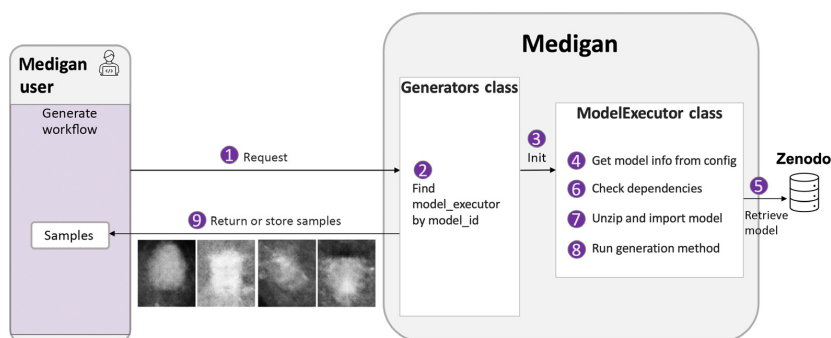


Fig. 5 The generated workflow. A user specifies a *model_id* in a request (1) to the generators class, which checks (2) if the model's *ModelExecutor* class instance is already initialized. If not, a new one is created (3), which (4) gets the model's config from the *global.json* dict, (5) loads the model (e.g., from *Zenodo*), (6) checks its dependencies, and (7) unzips and imports it, before running its internal generate function (8). Finally, the generated samples are returned to the user.

Code Snippet 2: Executing a *medigan* model for synthetic data generation.

```

1. from medigan import Generators

2. generators = Generators()

# create 100 polyps with masks using model 10 (FASTGAN)

generators.generate(model_id=10, num_samples=100)

```

3.5.2 Generate workflow extensions

Apart from storing or returning samples, a callable of the model's internal generate function can be returned to the user by setting *is_gen_function_returned*. This function with prepared but adjustable default arguments enables integration of the generate method into other workflows within *medigan* (e.g., model visualization) or outside of *medigan* (e.g., a user's AI model training). As a further alternative, a *torch*⁶⁷ dataset or dataloader can be returned for any model in *medigan* running *get_as_torch_dataset* or *get_as_torch_dataloader*, respectively. This further increases the versatility with which users can introduce *medigan*'s data synthesis capabilities into their AI model training and data preprocessing pipelines.

Instead of a user manually selecting a model via *model_id*, a model can also be automatically selected based on the recommendation from the model search and/or ranking methods. For instance, as triggered by Code Snippet 3, the models found in a search for *mammography* are ranked in ascending order based on FID, with the highest ranking model being selected and executed to generate the synthetic dataset.

3.6 Model Visualization

To allow users to explore the generative models in *medigan*, a novel model visualization module has been integrated into the library. It allows users to examine how changing inputs like the latent variable *z* and/or the class conditional label *y* (e.g., malignant/benign) can affect the generation process. Also, the correlation between multiple model outputs, such as the image and corresponding segmentation mask, can be observed and explored. Figure 6 illustrates an example showing an image-mask sample pair from *medigan*'s polyp generating FastGAN model.⁵¹

Code Snippet 3: Sequential searching, ranking, and data generation with highest ranked model.

```
1. from medigan import Generators
2. generators = Generators()
3. values = ['mammography'] # keywords for searching
4. metric = 'FID' # metric for ranking
5. generators.find_models_rank_and_generate
   (values=values, metric=metric)
```

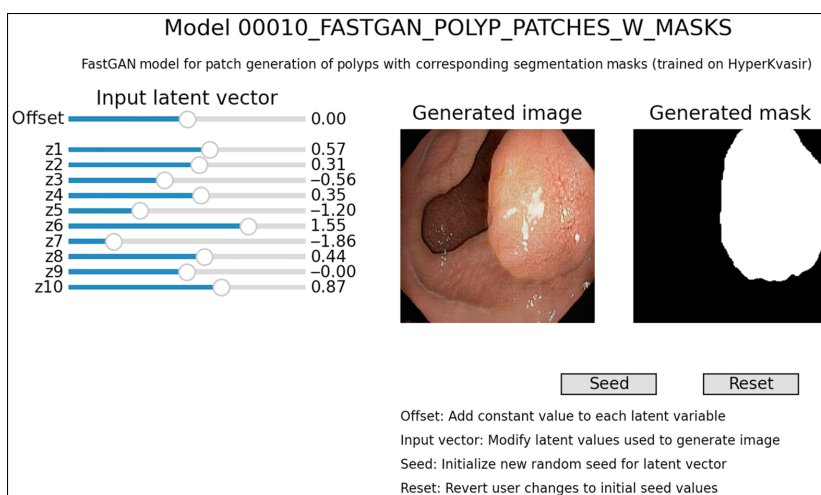


Fig. 6 Graphical user interface of *medigan*'s model visualization tool on the example of model 10, a FastGAN that synthesizes endoscopic polyp images with respective masks.⁵¹ The latent input vector can be adjusted via the sliders, reset via the *Reset* button, and sampled randomly via the *Seed* button.

This depiction of the graphical user interface (GUI) of the model visualization tool can be recreated by running Code Snippet 4.

Internally, the `model_visualizer` module retrieves a model's internal generate method as callable from the `model_executor` and adjusts the input parameters based on user interaction input from the GUI. This interaction further provides insight into a model's performance and capabilities. On one hand, it allows one to assess the fidelity of the generated samples. On the other hand, it also shows the model's captured sample diversity, i.e., as observed output variation over all possible input latent vectors. We leave the automation of manual visual analysis of this output variation to future work. For instance, such future work can use the `model_visualizer` to measure the variance of a reconstruction/perceptual error computed between pairs of images sampled from fixed-distance pairs of latent space vectors $\mathbf{z}$. The slider controls on the left of the interface allow one to change the latent variable, which for this specific model affects, for instance, polyp size, position, and background. As the size of the latent vector $\mathbf{z}$ commonly is relatively large, each n (e.g., 10) variables are grouped into one indexed slider resulting in z_m adjustable latent input variables. The seed button on the right allows one to initialize a new set of latent variables, which results in a new generated image. The reset buttons allows one to revert user's modifications to previous random values.

Code Snippet 4: Visualization of a model in *medigan*.

```

1. from medigan import Generators
2. generators = Generators()
3. n = 10 #grouping latent vector z dimensions by dividing them by 10
4. generators.visualize(model_id=10, slider_grouper=n)
# polyp with mask

```

3.7 Model Contribution

A core idea of *medigan* is to provide a platform where researchers can share and access trained models via a standardized interface. We provide in-depth instructions on how to contribute a model to *medigan* complemented by implementations automating parts of the model contribution process for users. In general, a pretrained model in *medigan* consists of a Python `__init__.py` and, in case the generation process is based on a machine learning model, a respective checkpoint or weights file. The former needs to contain a synthetic data storage method and a data generation method with a set of standardized parameters described in Sec. 3.5.1. Ideally, a model package further contains a license file, a *metadata.json* and/or a *requirements.txt* file, and a *test.sh* script to quickly verify the model's functionalities. To facilitate creation of these files, *medigan*'s GitHub repository provides model contributors with reusable templates for each of these files.

Keeping the effort of pretrained model inclusion to a minimum, the `generators` module contains a `contribute` function that initializes a `ModelContributor` class instance dedicated to automating the remainder of the model contribution process. This includes automated (i) validation of the user-provided *model_id*; (ii) validation of the path to the model's `__init__.py`; (iii) test of `importlib` import of the model as package; (iv) creation of the model's metadata dictionary; (v) adding the model metadata to *medigan*'s *global.json* metadata; (vi) end-to-end test of model with sample generation via `generators.test_model()`; (vii) upload of zipped model package to Zenodo via API; and (viii) creation of a GitHub issue, which contains the Zenodo link and model metadata, in the *medigan* repository. Being assigned to this GitHub issue, the *medigan* development team is notified about the new model, which can then be added via pull request. Code Snippet 5 shows how a user can run the `contribute` method illustrated in Fig. 7.

3.8 Model Testing Pipeline

Each new model contribution is being systematically tested before becoming part of *medigan*. For instance, on each submitted pull request to *medigan*'s GitHub repository, a CI pipeline automatically builds, formats, lints, and tests *medigan*'s codebase. This includes the automatic verification of each model's package, dependencies, compatibility with the interface, and correct functioning of its generation workflow. This allows one to ensure that all models and their metadata in the *global.json* file are available and working in a reproducible and standardized manner.

4 Applications

4.1 Community-Wide Data Access: Sharing the Essence of Restricted Datasets

medigan facilitates sharing and reusing trained generative models with the medical research community. On one hand, this reduces the need for researchers to retrain their own similar generative models, which can reduce the extensive carbon footprint⁹⁴ of deep learning in medical imaging. On the other hand, this provides a platform for researchers and data owners to share their dataset distribution without sharing the real data points of the dataset. Put differently,

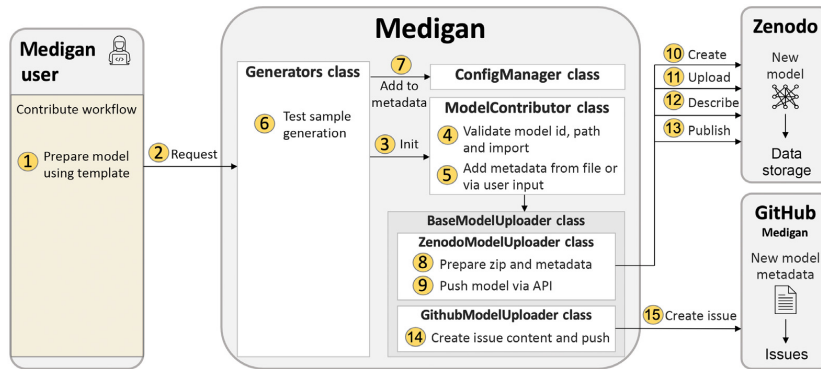


Fig. 7 Model contribution workflow. After model preparation (1), a user provides the model's id and metadata (2) to the generators class to (3) initialize a ModelContributor instance, which (4) validates and (5) extends the metadata. Next, (6) the model's sample generation capability is tested after (7) integration into *medigan*'s *global.json* model metadata. If successful, (8) the model package is prepared and (9–13) pushed to Zenodo via API. Lastly, (14 and 15) a GitHub issue containing the model metadata is created, assigned, and pushed to the *medigan* repository.

sharing generative models trained on (and instead of) patient datasets not only is beneficial as data curation step,¹⁴ but also minimizes the need to share images and personal data directly attributable to a patient. In particular, the latter can be quantifiably achieved when the generative model is trained using a differential privacy guarantee^{7,95} before being added to *medigan*. By reducing the barriers posed by data sharing restrictions and necessary patient privacy protection regulation, *medigan* unlocks a new paradigm of medical data sharing via generative models. This places *medigan* at the center toward solving the well-known issue of data scarcity^{7,9} in medical imaging.

Apart from that, *medigan*'s generative model contributors benefit from an increased exposure, dissemination, and impact of their work, as their generative models become readily usable by other researchers. As Table 3 illustrates, to date, *medigan* consists of 21 pretrained deep generative models contributed to the community. Among others, these include two conditional DCGAN models, six domain translation CycleGAN models and one mask-to-image pix2pix model. The training data comes from 10 different medical imaging datasets. Various of the

Code Snippet 5: Contribution of a model to *medigan*.

```

1. from medigan import Generators
2. generators = Generators()
3. generators.contribute(
4.     model_id = "00100_YOUR_MODEL", # assign ID
5.     init_py_path = "path/ending/with/__init__.py", # model package root
6.     generate_method_name = "generate", # method inside __init__.py
7.     model_weights_name = "10000",
8.     model_weights_extension = ".pt",
9.     dependencies = ["numpy", "torch"],
10.    zenodo_access_token = "TOKEN", #zenodo.org/account/settings/applications
11.    github_access_token = "TOKEN") #github.com/settings/tokens

```

models were trained on breast cancer datasets including INbreast,⁸¹ OPTIMAM,⁸² BCDR,⁸³ CBIS-DDSM,⁸⁶ and CSAW.⁸⁸ Models allow one to generate samples of different pixel resolutions ranging from regions-of-interest patches of size 128×128 and 256×256 to full images of 1024×1024 and 1332×800 pixels.

4.2 Investigating Synthetic Data Evaluation Methods

A further application of *medigan* is testing the properties of medical synthetic data. For instance, evaluation metrics for generative models can be readily tested in *medigan*'s multiorgan, multi-modality, and multimodel synthetic data setting.

Compared to generative modeling, synthetic data evaluation is a less explored research area.⁷ In particular, in medical imaging the existing evaluation frameworks, such as the FID⁴⁶ or the IS,¹⁷ are often limited in their applicability, as mentioned in Sec. 2.3. The models in *medigan* allow one to compare existing and new synthetic data evaluation metrics and their validation in the field of medical imaging. Multimodel synthetic data evaluation allows one to measure the correlation and statistical significance between synthetic data evaluation metrics and downstream task performance metrics. This enables the assessment of clinical usefulness of generative models on one hand and of synthetic data evaluation metrics on the other hand. In that sense, the metric itself can be evaluated including its variations when measured under different settings, datasets, or preprocessing techniques.

4.2.1 FID of *medigan* Models

We compute the FID to assess the models in *medigan* and report the results in Table 3. We further note that the FID can be computed not only between a synthetic and a real dataset (rs) but also between two sets of samples of the real dataset (rr). As the FID_{rr} describes the distance within two randomly sampled sets of the real data distribution, it can be used as an estimate of the real data variation and optimal lower bound for the FID_{rs} as shown in Table 3. Given the above, it follows that a high FID_{rr} likely also results in a higher FID_{rs} , which highlights the importance of accounting for the FID_{rr} when discussing the FID_{rs} . To do so, we propose the reporting of a FID ratio r_{FID} to describe the FID_{rs} in terms of the FID_{rr} .

$$r_{FID}(FID_{rs}, FID_{rr}) = 1 - \frac{FID_{rs} - FID_{rr}}{FID_{rs}}, \quad r_{FID} \in [0, 1] \subset \mathbb{R}. \quad (5)$$

Assuming $FID_{rs} \geq FID_{rr}$ bounds r_{FID} between 0 and 1, r_{FID} simplifies the comparison of FIDs computed using different models and datasets. A r_{FID} close to 1 indicates that much of the FID_{rs} can be explained by the general variation in the real dataset. The code used to compute the FID scores is available at <https://github.com/RichardObi/medigan/blob/main/tests/fid.py>.

The models in Table 3 yielding the highest ImageNet-based r_{FID} score are the ones with ID 10 (0.677, endoscopy, 256×256 , FastGAN), ID 13 (0.650, mammography, 1332×800 , CycleGAN), 14 (0.564, mammography, 1332×800 , CycleGAN), 20 (0.543, chest x-ray, 1024×1024 , PGGAN) and 1 (0.497, mammography, DCGAN, 128×128). This indicates that the r_{FID} does not depend on the modality, nor on the pixel resolution of the synthetic images. Further, neither image-to-image translation (e.g. CycleGAN) nor noise-to-image models (e.g., PGGAN, DCGAN, FastGAN) seem to have a particular advantage for achieving higher r_{FID} results.

The flow chart in Fig. 8 provides further insight into the comparison between the lower bound FID_{rr} and the model FID_{rs} . The red trend line shows a positive correlation between the FID_{rr} and FID_{rs} , which corroborates our previous assumption that a higher model FID_{rs} is to be expected given a higher lower bound FID_{rr} . Hence, for increased transparency, we motivate further studies to routinely report the lower bound FID_{rr} and the FID ratio r_{FID} apart from the model FID_{rs} . The three-channel RGB endoscopic images represented by orange dots have an FID_{rr} comparable with their grayscale radiologic counterparts. However, both chest x-ray datasets ChestX-ray14⁹⁰ and Node21⁸⁹ represented by green dots show a slightly lower FID_{rr} than other modalities. The model FID_{rs} shows a high variation across models without readily observable dependence on modality, generative model, or image size.

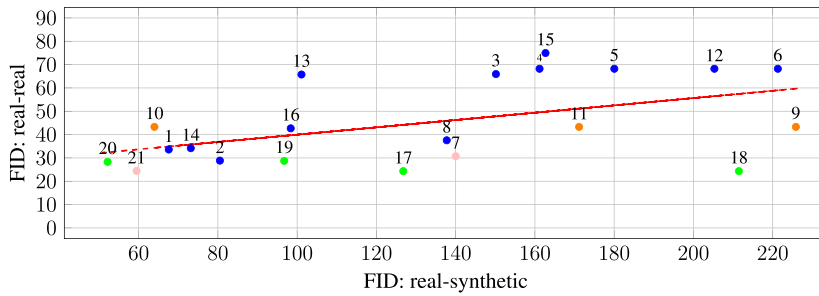


Fig. 8 Scatter plot illustrating the FID_{rs} of *medigan*'s models (real-synthetic) compared to the lower bound FID_r between two sets of the model's respective training dataset (real-real). The lower bound can represent an optimally achievable model and, as such, facilitates interpretation. Each model is represented by a dot below its model ID. The dots' color encoding depicts model modality, where blue: mammography, orange: endoscopy, green: chest x-ray, and pink: brain MRI. The red regression line illustrates the trend across all data points/models.

4.2.2 Analysing potential sources of bias in FID

The popular FID metric is computed based on the features of an Inception classifier (e.g., v1,⁴⁸ v3⁴⁷) trained on ImageNet⁵²—a database of natural images inherently different from the domain of medical images. This potentially limits the applicability of the FID to medical imaging data. Furthermore, the FID has been observed to vary based on the input image resizing methods and ImageNet backbone feature extraction model types.³¹ Based on this, we further hypothesize a susceptibility of the FID to variation due to (a) different backbone feature extractor weights and random seed initializations, (b) different medical and nonmedical backbone model pretraining datasets, (c) different image normalization procedures for real and synthetic dataset, (d) nuances between different frameworks and libraries used for FID calculation, and (f) the dataset sizes used to compute the FID.

Such variations can obstruct a reliable comparison of synthetic images generated by different generative models. Illustrating the potential of *medigan* to analyze such variations, we report and experiment with the FID. In particular, we subject the FID to variations in (i) the pretraining dataset of its backbone feature extractor and by (ii) testing the effects of image normalization across a set of *medigan* models. We experiment with the Inception v3 model trained on the recent RadImageNet dataset⁷² released as radiology-specific alternative to the ImageNet database.⁵² The RadImageNet-pretrained Inception v3 model weights we used are available at <https://github.com/BMEI-AI/RadImageNet>. We further compute the FID_{rs} and FID_r with and without normalization to analyze the respective impact on results.

In Table 4, the FID results are summarized allowing for cross-analysis between variations due to image normalization and/or due to the pretraining dataset of the FID feature extraction model. We observe generally lower FID values (1.15 to 7.32) for RadImageNet compared to ImageNet as FID model pretraining datasets (52.17 to 225.85). To increase FID comparability, we compute, as before, the FID ratio r_{FID} . The RadImageNet-based model results in notably lower r_{FID} values for both normalized and non-normalized images. Notably, an exception to this are models with ID 5 (mammography, 128×128 , DCGAN) and 6 (mammography, 128×128 , WGAN-GP) achieving respective RadImageNet-based r_{FID} scores of 0.593 and 0.550. In general, the RadImageNet-based model seems more robust at detecting if two sets of data originate from the same distribution resulting in low FID_r values. Overall, for most models, the FID is explained only by a limited amount by the variation in the real dataset and $r_{FID} < 0.7$ for all ImageNet and RadImageNet-based FIDs. The scatter plot in Fig. 9 further compares the RadImageNet-based FID with the ImageNet-FID for the models from Table 4. Noticeably, the difference between non-normalized and normalized images is surprisingly high for several models for both ImageNet and RadImageNet FIDs (e.g., models with IDs 6 and 8) while negligible for others (e.g., models with ID 1, 10, 13-16, and 19-21). Another observation is the relatively modest correlation between RadImageNet and ImageNet FID indicated by the slope of the red

Table 4 Normalized (left) and non-normalized (right) FID scores. This table measures the normalization impact on FID scores based on a promising set of *medigan*'s deep generative models. Synthetic samples were randomly-drawn for each model matching the number of available real samples. The *lower bound* FID_{rr} is computed between a pair of randomly sampled sets of real data (real-real), whereas the *model* FID_{rs} is computed between two randomly sampled sets of real and synthetic data (real-syn). The results for model 7 (Flair, T1, T1c, T2) and 21 (T1, T2) are averaged across modalities.

ID	Dataset	Normalized images						Non-Normalized images					
		FID _{ImageNet} ^{47,52}			FID _{RadImageNet} ⁷²			FID _{ImageNet} ^{47,52}			FID _{RadImageNet} ⁷²		
		real-real	real-syn	r_{FID}	real-real	real-syn	r_{FID}	real-real	real-syn	r_{FID}	real-real	real-syn	r_{FID}
1	Inbreast	33.61	67.60	0.497	0.25	1.27	0.197	28.59	66.76	0.428	0.29	1.15	0.252
2	Optimam	28.85	80.51	0.358	0.22	6.19	0.036	28.75	77.95	0.369	0.33	4.11	0.080
3	BCDR	65.94	150.16	0.439	0.80	3.00	0.265	66.25	149.33	0.444	0.80	3.10	0.259
5	BCDR	68.22	180.04	0.379	0.99	1.67	0.593	64.45	174.38	0.370	0.87	4.04	0.215
6	BCDR	68.22	221.30	0.308	0.99	1.80	0.550	64.45	206.57	0.312	0.87	2.95	0.295
7	BRATS 2018	30.73	140.02	0.219	0.07	5.31	0.012	30.73	144.00	0.215	0.07	6.53	0.010
8	CBIS-DDSM	37.56	137.75	0.272	0.46	3.05	0.151	32.06	91.09	0.352	0.36	6.58	0.055
10	HyperKvasir	43.31	63.99	0.677	0.11	7.32	0.015	43.31	64.01	0.677	0.11	7.33	0.015
12	BCDR	68.22	205.29	0.332	0.99	5.69	0.080	64.45	199.50	0.323	0.87	4.25	0.205
13	OPTIMAM	65.75	101.01	0.650	0.17	1.14	0.153	65.83	101.15	0.651	0.18	1.10	0.163
14	OPTIMAM	41.61	73.77	0.564	0.16	0.83	0.190	41.71	74.03	0.563	0.15	0.81	0.184
15	CSAW	74.96	162.67	0.461	0.31	4.07	0.076	73.62	165.53	0.445	0.19	3.71	0.051
16	CSAW	42.68	98.38	0.439	0.38	2.71	0.142	42.50	99.81	0.426	0.22	2.82	0.077
19	ChestX-ray14	28.75	96.74	0.297	0.19	0.77	0.243	28.75	96.78	0.297	0.19	0.66	0.286
20	ChestX-ray14	28.33	52.17	0.543	0.20	2.83	0.071	28.33	52.38	0.541	0.20	2.59	0.077
21	CrossMoDA	24.41	59.49	0.410	0.02	1.45	0.014	24.41	60.11	0.406	0.02	1.40	0.014

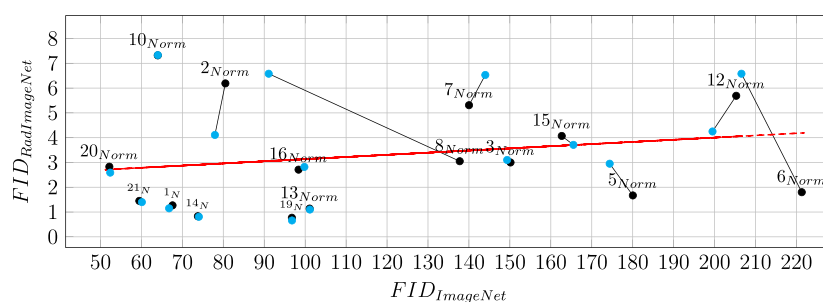


Fig. 9 Scatter plot demonstrating the FID_{rs} (real-synthetic) of *medigan* models from Table 4. The FID_{rs} is based on the features of two different inception classifiers,⁴⁷ one trained on ImageNet⁵² (x-axis) and the other trained on RadImageNet⁷² (y-axis). Each model is represented by a dot below its model ID. A black dot indicates an FID calculated from normalized (Norm/N) images, e.g., with pixel values scaled between 0 and 1, as opposed to a blue dot indicating an FID calculated from images without previous normalization. The dots that correspond to the same model IDs (normalized and non-normalized) are connected via black lines. The red regression line illustrates the trend across all data points.

regression line. Counterexamples for this correlation include model 2 (normalized), which has a low ImageNet-based FID (80.51) compared to a high RadImageNet-based FID (6.19), and model 6 (normalized), which, in contrast, has a high ImageNet-based FID (221.30) and a low RadImageNet-based FID (1.80). With a low ImageNet-based FID (63.99), but surprisingly high RadImageNet-based FID (7.32), model 10 (both normalized and non-normalized) is a further counterexample. The example of model 10 is of particular interest, as it indicates limited applicability of the Radiology-specific RadImageNet-based FID for out-of-domain data, such as three-channel endoscopic images.

Given the demonstrated high impact of backbone model training set and image normalization on FID, it is to be recommended that studies specify the exact model used for FID calculation and any applied data preprocessing and normalization steps. Further, where possible, reporting the RadImageNet-based FID allows for reporting a radiology domain-specific FID. The latter is seemingly less susceptible to variation in the real datasets than the ImageNet-based FID while also being capable of capturing other, potentially complementary, patterns in the data.

4.3 Improving Clinical Medical Image Analysis

A high-impact clinical application of synthetic data is the improvement of clinical downstream task performance such as classification, detection, segmentation, or treatment response estimation. This can be achieved by using image synthesis for data augmentation, domain adaptation, and data curation (e.g., artifact removal, noise reduction, super-resolution)^{7,63} to enhance the performance of clinical decision support systems such as computer-aided diagnosis (CADx) and detection (CAdE) software.

In Table 5, the capability of improving the clinical downstream task performance is demonstrated for various *medigan* models and modalities. Downstream task models trained on a combination of real and synthetic imaging data achieve promising results surpassing the alternative results achieved from training only on real data. The results are taken from the respective publications^{11,14,50,84} and indicate that image synthesis can further improve the promising performance demonstrated by deep learning-based CADx and CAdE systems, e.g., in mammography⁹⁶ and brain MRI.⁸⁵ For downstream task evaluation, we generally note the importance of avoiding data leakage between training, validation, and test sets by training the

Table 5 Examples of the impact of synthetic data generated by *medigan* models on downstream task performance. Based on real test data, we compare the performance metrics of a model trained *only on real* data with a model trained on real data *augmented with synthetic* data. The metrics are taken from the respective publications describing the models.

ID	Test dataset	Task	Metric	Trained on real	Real + synthetic
2	OPTIMAM	Mammogram patch classification ¹¹	F1	0.90	0.96
3	BCDR	Mammogram mass detection ⁵⁰	FROC AUC	0.83	0.89
5	BCDR	Mammogram patch classification ¹⁴	F1	0.891	0.920
5	BCDR	Mammogram patch classification ¹⁴	AUROC	0.928	0.959
5	BCDR	Mammogram patch classification ¹⁴	AUPRC	0.986	0.992
6	BCDR	Mammogram patch classification ¹⁴	F1	0.891	0.969
6	BCDR	Mammogram patch classification ¹⁴	AUROC	0.928	0.978
6	BCDR	Mammogram patch classification ¹⁴	AUPRC	0.986	0.996
7	BRATS 2018	Brain tumor segmentation ⁸⁴	Dice	0.796	0.814
14	OPTIMAM	Mammogram mass detection ⁵⁰	FROC AUC	0.83	0.85
15	OPTIMAM	Mammogram mass detection ⁵⁰	FROC AUC	0.83	0.85

Table 6 Examples of the impact of synthetic data generated by *medigan* models on downstream task performance. Based on real test data, we compare the performance metrics of a model trained *only on real* data with a model trained *only on synthetic* data. The metrics are taken from the respective publications describing the models. *n.a.* refers to the case where only synthetic data can be used, as no annotated real training data is available.

ID	Test dataset	Task	Metric	Trained on real	Trained on synthetic
4	BCDR	Mammogram mass segmentation	Dice	0.865	0.737
11	HyperKvasir	Polyp segmentation ⁵¹	Dice loss	0.112	0.137
11	HyperKvasir	Polyp segmentation ⁵¹	IoU	0.827	0.798
11	HyperKvasir	Polyp segmentation ⁵¹	F-Score	0.888	0.863
20	ChestX-ray14	Lung disease classification ⁹¹	AUROC	0.947	0.878
21	CrossMoDA	Brain tumor segmentation ⁹²	Dice	n.a.	0.712
21	CrossMoDA	Cochlea segmentation ⁹²	Dice	n.a.	0.478

generative model either on only the dataset partition used to train the respective downstream task model (e.g., IDs 2, 3, 7, 14, 15) or to train the generative models on an entirely different dataset (e.g., IDs 5, 6).

The approaches displayed in Table 6 represent the application, where synthetic data is used instead of real data to train downstream task models. Despite an observable performance decrease when training on synthetic data only, the results^{51,91,92} demonstrate the usefulness of synthetic data if none or only limited real training data is available or shareable. For example, if labels or annotations in a target domain are scarce but present in a source domain, a generative model can translate annotated data from the source domain to the target domain to enable supervised training of downstream task models.^{92,93}

5 Discussion and Future Work

In this work, we introduced *medigan*, an open-source Python library, which allows one to share pretrained generative models for synthetic medical image generation. The package is easily integrable into other packages and tools, including commercial ones. Synthetic data can enhance the performance, capabilities, and robustness of data-hungry deep learning models as well as to mitigate common issues such as domain shift, data scarcity, class imbalance, and data privacy restrictions. Training one's own generative network can be complex and expensive since it requires a considerable amount of time, effort, specific dedicated hardware, carbon emissions, as well as knowledge and applied skills in generative AI. An alternative and complementary solution is the distribution of pretrained generative models to allow their reuse by AI researchers and engineers worldwide.

medigan can help to reduce the time to run synthetic data experiments and can readily be added as a component, e.g., as a dataloader as discussed in Sec. 3.5.2, in AI training pipelines. As such, the generated data can be used to improve supervised learning models as described in Sec. 4.3 via training or fine-tuning but can also serve as plug-and-play data source for self/semisupervised learning, e.g., to pretrain clinical downstream task models.

Furthermore, studies that use additional synthetic training data for training deep learning models often do not report all the specifics about their underlying generative model.^{7,75} Within *medigan*, each generative model is documented, openly accessible, and reusable. This increases the reproducibility of studies that use synthetic data and makes it more transparent where the data or parts thereof originated from. This can help to achieve the traceability objectives outlined in the FUTURE-AI consensus guiding principles toward AI trustworthiness in medical imaging.⁷⁵ *medigan*'s currently 21 generative models are illustrated in Table 3 and developed and validated by AI researchers and/or specialized medical doctors. Furthermore, each model contains traceable⁷⁵ and version-controlled metadata in *medigan*'s *global.json* file, as outlined in

Sec. 3.3. The searchable (see Sec. 3.4) metadata allows one to choose a suitable model for a user's task at hand and includes, among others, the dataset used during the training process, the trained date, publication, modality, input arguments, model types, and comparable evaluation metrics.

To assess model suitability, users are recommended to first (i) ensure the compatibility between their planned downstream task (e.g., mammogram region-of-interest classification) and a candidate *medigan* model (e.g., mammogram region-of-interest generator). Second, (ii) a user's real (test) data and the model's synthetic data should be compatible corresponding, for instance, in domain, organ, or disease manifestation. If the awareness of the domain shifts between real and synthetic data remains limited after this qualitative analysis, (iii) a quantitative assessment (e.g., via FID) is recommended. Finally, (iv) it is to be assessed if a downstream task improvement is plausible. This depends, among others, on the tested scenario and the task at hand, but also on the amount, domain, task specificity and quality of the available real data, and the generative model's capabilities as indicated by its reported evaluation metrics from previous studies. If a positive impact of synthetic data on downstream task performance is plausible, users are recommended to proceed toward empirical verification.

The exploration and multimodel evaluation of the properties of generative models and synthetic data is a further application of *medigan*. *medigan*'s visualization tool (see Sec. 3.6) intuitively allows the user to explore and adjust the input latent vector of generative models to visually evaluate, e.g., its inherent diversity and condition adherence⁷ (i.e., how well does a given mask or label fit the generated image). The evaluation of synthetic data by human experts, such as radiologists, is a costly and time-consuming task, which motivates the usage of automated metric-based evaluation such as the FID. Our multimodel analysis reveals sources of bias in FID reporting. We show the susceptibility of FID to vary substantially based on changes in input image normalization or in the choice of the pretraining dataset of the FID feature extractor. This finding highlights the need to report the specific models, preprocessing, and implementations used to compute the FID alongside the FID ratio r_{FID} proposed in Sec. 4.2.1 to account for the variation immanent in the real dataset. With *medigan* model experiments demonstrably leading to insights in synthetic data evaluation, future research can use *medigan* as a tool to accelerate, test, analyze, and compare new synthetic data and generative model evaluation and exploration techniques.

5.1 Legal Frameworks for Sharing of Synthetic and Real Patient Data

Many countries have enacted regulations that govern the use and sharing of data related to individuals. The two most recognized legal frameworks are the Health Insurance Portability and Accountability Act (HIPAA)⁹⁷ from the United States (U.S.) and the General Data Protection Regulation (GDPR)⁹⁸ from the European Union (E.U.). These regulations govern the use and disclosure of individuals' protected health information (PHI) and assures individuals' data is protected while allowing use for providing quality patient care.^{99–102}

Conceptually, synthetic data is not real data about any particular individual and conversely to real data, synthetic data can be generated at high volumes and potentially shared without restriction. In this sense, under both GDPR and HIPAA regulation, the rules govern the use of real data for the generation and evaluation of synthetic datasets, as well as the sharing of the original dataset. However, once fully synthetic data is generated, this new dataset falls outside the scope of the current regulations based on the argument that there is no direct correlation between the original subjects and the synthetic subjects. A common interpretation is that as long as the real data remains in a secure environment during the generation of synthetic data, there is little to no risk to the original subjects.¹⁰³

As a consequence, the use of synthetic data can help prevent researchers from inadvertently using and possibly exposing patients identifiable data. Synthetic data can also lessen the controls imposed by Institutional Review Boards (IRBs) and based on international regulations by ensuring data is never mapped to real individuals.¹⁰⁴ There are multiple methods of generating synthetic data, some of which include building models from real data, which can create a set statistically similar to real data. How similar the synthetic data is to real-world data often defines its "utility," which will vary depending on the synthesis methods used and the needs of the study at hand. If the utility of the synthetic data is high enough then evaluation results are expected to

be similar to those that use real data.¹⁰³ Being built based on real data, a common concern is patient reidentification and leaking of patient-specific features in generative models.^{7,15} Despite the arguably permissive aforementioned regulations, deidentification⁶³ of the training data prior to generative model training is to be recommended. This can minimize the possibility of generative models leaking sensitive patient data during inference and after sharing. A further recommended and mathematically-proven tool for privacy preservation is differential privacy (DP).⁹⁵ DP can be included in the training of deep generative model, among other setups, by adding DP noise to the gradients.

5.2 Expansion of Available Models

In the future, further generative models across medical imaging disciplines, modalities, and organs can be integrated into *medigan*. The capabilities of additional models can range from privatising or translating the user's data from one domain to another, balancing or debiasing imbalanced datasets, reconstructing, denoising or removing artifacts in medical images, or resizing images, e.g., using image super-resolution techniques. Despite *medigan*'s current focus on models based on GANs,¹⁶ the inclusion of different additional types of generative models is desirable and will enable insightful comparisons. In particular, this is to be further emphasized considering the recent successes of diffusion models,^{25–27} variational autoencoders,²¹ and normalizing flows^{22–24} in the computer vision and medical imaging^{105–107} domains. Before integrating and testing a new model via the pipeline described in Sec. 3.8, we assess whether a model is to become a candidate for inclusion into *medigan*. This threefold assessment is based on the SynTRUST framework⁷ and reviews whether (1) the model is well-documented (e.g., in a respective publication), (2) the model or its synthetic data is applicable to a task of clinical relevance, and (3) whether the model has been methodically validated.

5.3 Synthetic DICOM Generation

Since the dominant data format used for medical imaging is Digital Imaging and Communications in Medicine (DICOM), we plan to enhance *medigan* by integrating the generation of DICOM compliant files. DICOM consists of two main components, pixel data and the DICOM header. The latter can be described as an embedded dataset rich with information related to the pixel data such as the image sequence, patient, physicians, institutions, treatments, observations, and equipment.⁶³ Future work will explore combining our GAN generated images with synthetic DICOM headers. The latter will be created from the same training images from which the *medigan* models are trained to create synthetic DICOM data with high statistical similarity to real-world data. In this regard, a key research focus will be the creation of an appropriate and DICOM-compliant description of the image acquisition protocol for a synthetic image. The design and development of an open-source software package for generating DICOM files based on synthesized DICOM headers associated to (synthetic) images will extend prior work¹⁰⁸ that demonstrated the generation of synthetic headers for the purpose of evaluating deidentification methods.

6 Conclusion

We presented the open-source *medigan* package, which helps research in medical imaging to rapidly create synthetic datasets for a multitude of purposes such as AI model training and benchmarking, data augmentation, domain adaptation, and intercentre data sharing. *medigan* provides simple functions and interfaces for users, allowing one to automate generative model search, ranking, synthetic data generation, and model contribution. By reuse and dissemination of existing generative models in the medical imaging community, *medigan* allows researchers to speed up their experiments with synthetic data in a reproducible and transparent manner.

We discuss three key applications of *medigan*, which include (i) sharing of restricted datasets, (ii) improving clinical downstream task performance, and (iii) analyzing the properties of generative models, synthetic data, and associated evaluation metrics. Ultimately, the aim of *medigan* is to contribute to benefiting patients and clinicians, e.g., by increasing the performance and robustness of AI models in clinical decision support systems.

Disclosures

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Acknowledgments

We would like to thank all model contributors, such as Alyafi et al.,¹¹ Szafranowska et al.,¹⁴ Thambawita et al.,⁵¹ Kim et al.,⁸⁴ Segal et al.,⁹¹ Joshi et al.,⁹² and Garrucho et al.⁵⁰ This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No. 952103 and No. 101057699. Eloy García and Kaisar Kushibar hold the Juan de la Cierva fellowship from the Ministry of Science and Innovation of Spain with reference numbers FJC2019-040039-I and FJC2021-047659-I, respectively.

Data, Materials, and Code Availability

medigan is a free Python (v3.6+) package published under the MIT license and distributed via the Python Package Index (<https://pypi.org/project/medigan/>). The package is open-source and invites the community to contribute on GitHub (<https://github.com/RichardObi/medigan>). A detailed documentation of *medigan* is available (<https://medigan.readthedocs.io/en/latest/>) that contains installation instructions, the API reference, a general description, code examples, a testing guide, a model contribution user guide, and documentation of the generative models available in *medigan*.

References

1. C. Martin-Isla et al., "Image-based cardiac diagnosis with machine learning: a review," *Front. Cardiovasc. Med.* **7**, 1 (2020).
2. R. Aggarwal et al., "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," *NPJ Digital Med.* **4**(1), 1–23 (2021).
3. X. Liu et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *Lancet Digital Health* **1**(6), e271–e297 (2019).
4. J. Schlemper et al., "A deep cascade of convolutional neural networks for MR image reconstruction," *Lect. Notes Comput. Sci.* **10265**, 647–658 (2017).
5. E. Ahishakiye et al., "A survey on deep learning in medical image reconstruction," *Intell. Med.* **1**(03), 118–127 (2021).
6. N. Tajbakhsh et al., "Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation," *Med. Image Anal.* **63**, 101693 (2020).
7. R. Osuala et al., "Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging," *Med. Image Anal.* **84**, 102704 (2023).
8. C. Jin et al., "Predicting treatment response from longitudinal images using multi-task deep learning," *Nat. Commun.* **12**, 1851 (2021).
9. W. L. Bi et al., "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: Cancer J. Clin.* **69**(2), 127–157 (2019).
10. F. Prior et al., "Open access image repositories: high-quality data to enable machine learning research," *Clin. Radiol.* **75**(1), 7–12 (2020).
11. B. Alyafi, O. Diaz, and R. Marti, "DCGANs for realistic breast mass augmentation in x-ray mammography," *Proc. SPIE* **11314**, 1131420 (2020).
12. X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: a review," *Med. Image Anal.* **58**, 101552 (2019).
13. J. M. Wolterink et al., "Deep MR to CT synthesis using unpaired data," *Lect. Notes Comput. Sci.* **10557**, 14–23 (2017).
14. Z. Szafranowska et al., "Sharing generative models instead of private data: a simulation study on mammography patch classification," *Proc. SPIE* **12286**, 122860Q (2022).
15. T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data—anonymisation groundhog day," in *31st USENIX Secur. Symp. (USENIX Security 22)*, pp. 1451–1468 (2022).

16. I. Goodfellow et al., “Generative adversarial nets,” in *Adv. Neural Inf. Process. Syst.*, pp. 2672–2680 (2014).
17. T. Salimans et al., “Improved techniques for training GANs,” in *Adv. Neural Inf. Process. Syst.* 29, pp. 2234–2242 (2016).
18. L. Mescheder, A. Geiger, and S. Nowozin, “Which training methods for GANs do actually converge?,” in *Int. Conf. Mach. Learn.*, PMLR, pp. 3481–3490 (2018).
19. S. Arora, A. Risteski, and Y. Zhang, “Do GANs learn the distribution? Some theory and empirics,” in *Int. Conf. Learn. Represent.* (2018).
20. L. Ruthotto and E. Haber, “An introduction to deep generative modeling,” *GAMM-Mitteilungen* 44(2), e202100008 (2021).
21. D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” arXiv:1312.6114 (2013).
22. D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *Int. Conf. Mach. Learn.*, PMLR, pp. 1530–1538 (2015).
23. L. Dinh, D. Krueger, and Y. Bengio, “Nice: non-linear independent components estimation,” arXiv:1410.8516 (2014).
24. L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real NVP,” arXiv:1605.08803 (2016).
25. J. Sohl-Dickstein et al., “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Int. Conf. Mach. Learn.*, PMLR, pp. 2256–2265 (2015).
26. Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” in *Adv. Neural Inf. Process. Syst.* 32 (2019).
27. J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Adv. Neural Inf. Process. Syst.* 33, pp. 6840–6851 (2020).
28. M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Int. Conf. Mach. Learn.*, PMLR, pp. 214–223 (2017).
29. I. Gulrajani et al., “Improved training of wasserstein gans,” arXiv:1704.00028 (2017).
30. B. Liu et al., “Towards faster and stabilized gan training for high-fidelity few-shot image synthesis,” in *Int. Conf. Learn. Represent.* (2020).
31. M. Kang, J. Shin, and J. Park, “StudioGAN: a taxonomy and benchmark of GANs for image synthesis,” arXiv:2206.09479 (2022).
32. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv:1511.06434 (2015).
33. T. Karras et al., “Progressive growing of gans for improved quality, stability, and variation,” arXiv:1710.10196 (2017).
34. M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv:1411.1784 (2014).
35. A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier GANs,” in *Int. Conf. Mach. Learn.*, PMLR, pp. 2642–2651 (2017).
36. P. Isola et al., “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1125–1134 (2017).
37. J.-Y. Zhu et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
38. Y. Choi et al., “Stargan: unified generative adversarial networks for multi-domain image-to-image translation,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 8789–8797 (2018).
39. T. Park et al., “Semantic image synthesis with spatially-adaptive normalization,” in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 2337–2346 (2019).
40. V. Sushko et al., “OASIS: only adversarial supervision for semantic image synthesis,” *Int. J. Comput. Vision* 130(12), 2903–2923 (2022).
41. T. R. Shaham, T. Dekel, and T. Michaeli, “Singan: learning a generative model from a single natural image,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 4570–4580 (2019).
42. D. Korkinof et al., “Perceived realism of high resolution generative adversarial network derived synthetic mammograms,” *Radiol.: Artif. Intell.* 3, e190181 (2020).
43. B. Alyafi et al., “Quality analysis of DCGAN-generated mammography lesions,” *Proc. SPIE* 11513, 115130B (2020).
44. A. Borji, “Pros and cons of GAN evaluation measures,” *Comput. Vision Image Understanding* 179, 41–65 (2019).

45. A. Borji, "Pros and cons of GAN evaluation measures: new developments," *Comput. Vision Image Understanding* **215**, 103329 (2022).
46. M. Heusel et al., "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inf. Process. Syst.* **30** (2017).
47. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2818–2826 (2016).
48. C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1–9 (2015).
49. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
50. L. Garrucho et al., "High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection," *Front. Oncol.* **12**, 1044496 (2022).
51. V. Thambawita et al., "SinGAN-Seg: synthetic training data generation for medical image segmentation," *PLoS One* **17**(5), e0267976 (2022).
52. J. Deng et al., "Imagenet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 248–255 (2009).
53. accel brain, "Generative adversarial networks library: Pygan," 2021, <https://github.com/accel-brain/accel-brain-code/tree/master/Generative-Adversarial-Networks/>.
54. A. Pal and A. Das, "TorchGAN: a flexible framework for gan training and evaluation," *J. Open Source Software* **6**(66), 2606 (2021).
55. J. Herzer and T. NeuerRadu, "vegans," 2021, <https://github.com/unit8co/vegans/>.
56. NVIDIA, "Imaginaire," 2021, <https://github.com/NVlabs/imaginaire>.
57. J. Shor, "Tensorflow-GAN (TF-GAN): tooling for gans in tensorflow," 2022, <https://github.com/tensorflow/gan>.
58. E. Linder-Norén, "Keras-GAN: Pytorch implementations of generative adversarial networks," 2021, <https://github.com/eriklindernoren/PyTorch-GAN>.
59. E. Linder-Norén, "Keras-GAN: Keras implementations of generative adversarial networks," 2022, <https://github.com/eriklindernoren/Keras-GAN>.
60. K. S. Lee and C. Town, "Mimicry: towards the reproducibility of GAN research," arXiv:2005.02494 (2020).
61. T. Wolf et al., "Transformers: state-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods in Nat. Language Process.: Syst. Demonstrations*, pp. 38–45 (2020, October).
62. M. Kahng et al., "GAN lab: understanding complex deep generative models using interactive visual experimentation," *IEEE Trans. Vision Comput. Graphics* **25**(1), 310–320 (2018).
63. O. Diaz et al., "Data preparation for artificial intelligence in medical imaging: a comprehensive guide to open-access platforms and tools," *Phys. Med.* **83**, 25–37 (2021).
64. F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Methods Prog. Biomed.* **208**, 106236 (2021).
65. C. M. Moore et al., "CleanX: a Python library for data cleaning of large sets of radiology images," *J. Open Source Software* **7**(76), 3632 (2022).
66. M. J. Cardoso et al., "MONAI: an open-source framework for deep learning in healthcare," arXiv:2211.02701 (2022).
67. A. Paszke et al., "Pytorch: an imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst.* **32**, H. Wallach et al., Eds., pp. 8024–8035, Curran Associates, Inc. (2019).
68. E. Gibson et al., "Niftynet: a deep-learning platform for medical imaging," *Comput. Methods Prog. Biomed.* **158**, 113–122 (2018).
69. N. Pawlowski et al., "DLTK: state of the art reference implementations for deep learning on medical images," arXiv:1711.06853 (2017).
70. A. Nikolaos, "Deep learning in medical image analysis: a comparative analysis of multi-modal brain-MRI segmentation with 3D deep neural networks," Master's thesis, University of Patras (2019).

71. M. Baumgartner et al., “nnDetection: a self-configuring method for medical object detection,” *Lect. Notes Comput. Sci.* **12905**, 530–539 (2021).
72. X. Mei et al., “RadImageNet: an open radiologic deep learning research dataset for effective transfer learning,” *Radiol.: Artif. Intell.* **4**, e210315 (2022).
73. The Python Package Index, “medigan 1.0.0,” 2022, <https://pypi.org/project/medigan/> (accessed 5 February 2023).
74. R. Osuala, G. Skorupko, and N. Lazrak, “medigan getting started,” 2022, <https://medigan.readthedocs.io/en/latest> (accessed 5 February 2023).
75. K. Lekadir et al., “FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging,” arXiv:2109.09658 (2021).
76. EuCanImage Consortium, “EuCanImage towards a European cancer imaging platform for enhanced artificial intelligence in oncology,” 2020, <https://eucanimage.eu/> (accessed 5 February 2023).
77. M. Abadi et al., “TensorFlow: large-scale machine learning on heterogeneous systems,” 2015, tensorflow.org.
78. F. Chollet et al., “Keras,” 2015, <https://github.com/fchollet/keras>.
79. C. Larman, *Applying UML and Pattern: An Introduction to Object Oriented Analysis and Design and the Unified Process*, Prentice Hall PTR (2001).
80. E. Gamma et al., *Design Patterns: Elements of Reusable Object-Oriented Software*, Pearson Deutschland GmbH (1995).
81. I. C. Moreira et al., “INbreast: toward a full-field digital mammographic database,” *Acad. Radiol.* **19**(2), 236–248 (2012).
82. M. D. Halling-Brown et al., “Optimam mammography image database: a large-scale resource of mammography images and clinical data,” *Radiol.: Artif. Intell.* **3**, e200103 (2020).
83. M. G. Lopez et al., “BCDR: a breast cancer digital repository,” in *15th Int. Conf. Exp. Mech.*, Vol. 1215 (2012).
84. S. Kim, B. Kim, and H. Park, “Synthesis of brain tumor multicontrast MR images for improved data augmentation,” *Med. Phys.* **48**(5), 2185–2198 (2021).
85. B. H. Menze et al., “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2014).
86. R. S. Lee et al., “A curated mammography data set for use in computer-aided detection and diagnosis research,” *Sci. Data* **4**(1), 170177 (2017).
87. H. Borgli et al., “Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Sci. Data* **7**(1), 283 (2020).
88. K. Dembrower, P. Lindholm, and F. Strand, “A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (CSAW),” *J. Digital Imaging* **33**(2), 408–413 (2020).
89. E. Sogancioglu, K. Murphy, and B. van Ginneken, “NODE21 (v-5) [data set],” *Zenodo* (2021).
90. X. Wang et al., “Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2097–2106 (2017).
91. B. Segal et al., “Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing GANs,” *SN Comput. Sci.* **2**(4), 1–17 (2021).
92. S. Joshi et al., “nn-UNet training on CycleGAN-translated images for cross-modal domain adaptation in biomedical imaging,” *Lect. Notes Comput. Sci.* **12963**, 540–551 (2022).
93. R. Dorent et al., “CrossMoDA 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation,” *Med. Image Anal.* **83**, 102628 (2022).
94. R. Selvan et al., “Carbon footprint of selecting and training deep learning models for medical image analysis,” in *Med. Image Comput. and Comput. Assist. Intervention—MICCAI 2022: 25th Int. Conf.*, 18–22 September 2022, Singapore, Proceedings, Part V, pp. 506–516, Springer Nature, Cham, Switzerland (2022).
95. C. Dwork et al., “The algorithmic foundations of differential privacy,” *Found. Trends® Theor. Comput. Sci.* **9**(3–4), 211–407 (2014).

96. L. Abdelrahman et al., “Convolutional neural networks for breast cancer detection in mammography: a survey,” *Comput. Biol. Med.* **131**(Jan.), 104248 (2021).
97. Centers for Medicare & Medicaid Services, “The Health Insurance Portability and Accountability Act of 1996 (HIPAA),” 1996, <http://www.cms.hhs.gov/hipaa/>.
98. European Parliament and Council of European Union, “Council regulation (EU) no 2016/679,” 2018, <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679/>.
99. Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule, Board on Health Sciences Policy, Board on Health Care Services et al., *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*, p. 12458, National Academies Press, Washington, D.C. (2009).
100. U.S. Dept. of Health and Human Services, “Summary of the HIPAA privacy rule: HIPAA compliance assistance,” 2003, <http://purl.fdlp.gov/GPO/gpo9756>.
101. S. M. Shah and R. A. Khan, “Secondary use of electronic health record: opportunities and challenges,” *IEEE Access* **8**, 136947–136965 (2020).
102. C. F. Mondschein and C. Monda, “The EU’s general data protection regulation (GDPR) in a research context,” in *Fundamentals of Clinical Data Science*, P. Kubben, M. Dumontier, and A. Dekker, Eds., pp. 55–71, Springer International Publishing, Cham (2019).
103. K. El Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*, 1st ed., O’Reilly Media, Inc, Sebastopol, California (2020).
104. F. K. Dankar and M. Ibrahim, “Fake it till you make it: guidelines for effective synthetic data generation,” *Applied Sciences* **11**, 2158 (2021).
105. W. H. L. Pinaya et al., “Brain imaging generation with latent diffusion models,” in *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 117–126, Springer Nature, Cham, Switzerland (2022, October).
106. W. H. L. Pinaya et al., “Unsupervised brain imaging 3D anomaly detection and segmentation with transformers,” *Med. Image Anal.* **79**, 102475 (2022).
107. N. Pawlowski, D. Coelho de Castro, and B. Glocker, “Deep structural causal models for tractable counterfactual inference,” in *Adv. Neural Inf. Process. Syst.* **33**, pp. 857–869 (2020).
108. M. Rutherford et al., “A DICOM dataset for evaluation of medical image de-identification,” *Sci. Data* **8**, 183 (2021).

Biographies of the authors are not available.

Chapter 7

Results and Discussion

This thesis represents a comprehensive exploration of the challenges and advancements in leveraging generative models for cancer imaging, with a focus on addressing critical issues such as data scarcity, privacy, domain adaptation, and clinical integration. Throughout the work, novel frameworks and methodologies were proposed, validated, and applied to enhance tasks like synthetic data generation, tumor localization, diagnostic accuracy, and privacy-preserving machine learning. In this chapter, we provide a reflective account of the methods, applications and findings presented in the preceding chapters, alongside considering their limitations, implications and potential for future research.

7.1 Contrast Localization and Characterization

7.1.1 Modeling of Contrast Patterns

In Chapter 3 and Chapter 4, deep generative models were introduced for synthesizing how non-contrast-enhanced MRI scans would visually appear if contrast media would have been intravenously injected before acquisition of the MRI scan. The network architectures presented for this task are (a) a conditional generative adversarial network in Chapter 3 and (b) a multi-conditional ControlNet latent diffusion model in Chapter 4. The application of such generative models to this task introduces a non-invasive, cost-effective, and clinically valuable alternative to the traditional injection of gadolinium-based contrast agents (GBCA). The generative models were able to learn temporal patterns based on learning from multiple DCE-MRI sequences acquired in irregular

7.1. Contrast Localization and Characterization

time intervals after contrast agent injection. The respective time-signal intensity curves observed and modelled in these sequences reflect the uptake, wash-in, peak contrast, and wash-out phases of the contrast agent over time [227, 61, 117]. Observing these temporal patterns in the tumor area enables tumor characterization such as tumor subtype and malignancy determination, and treatment response prediction and planning [30, 208, 35]. Apart from tumor characterization, generative models are also useful for tumor diagnosis and treatment planning. This is exemplified by their capability, after learning the task of predicting contrast in non-contrast enhanced MRI, to localize, detect and potentially also segment tumors without the need of direct supervision via segmentation masks.

7.1.2 Impact and Benefit of Contrast Prediction Applications

Detecting tumors via virtual contrast injection is particularly relevant considering the existence of multiple patient populations which receive breast MRI cancer screening examinations in breast cancer screening programs. This is common, for example, in patients with harmful BRCA1 or BRCA2 genetic changes, which have a prevalence ranging between 0.2% to 0.3% in the general population [161]. Apart from unlocking new diagnostic options, the simulation of contrast-enhancement does not bear the risks associated with contrast injections, benefiting, among others, patients with contrast agent contraindications such as pregnancy or kidney diseases. Simulating contrast as opposed to using only non-contrast enhanced scans, can unlock better diagnostic and therapeutic capabilities for these patients. One example are patients with chronic kidney disease, which, as a progressive condition, has a global prevalence between 11% to 13%, amounting to over 800 million individuals [90, 114]. In terms of adverse health effects after iodinated contrast media (ICM) administration, overall incidence rates have been reported to range between 0.15% and 0.69%, with severe reactions occurring in 0.005% to 0.013% of cases, as assessed in a retrospective review of 456,930 doses in a single center [95]. The study defined adverse effects as a reaction occurring in the radiology suite during contrast administration or within 30 minutes of administration [95]. It is to be noted that the reported numbers refer to a patient population that had been previously filtered, as contrast agents were likely not administered in patients with known allergies or any other contraindications. With some patients further receiving premedication, the aforementioned percentages likely underrepresent the true risk of adverse reactions in a broader unfiltered population. Apart from that, also the long-term effects of contrast agents, such as gadolinium deposition in the body,

are to be taken into consideration in separate long-term studies and assessments of correlation and causality with health-related observations [165, 98, 165]. In this regard, by simulating contrast uptake over multiple timepoints, generative models can unlock personalized treatment planning for large patient cohorts, while also reducing imaging costs, scan times, and patient discomfort, coupled with reduced examination times for patients, reduction of motion artifacts, and reduced risk of adverse health effects. Additionally, the generated synthetic data, as additional training or inference data, can improve machine learning models, e.g. for tumor classification, detection, and segmentation. Furthermore, the aforementioned generative models can become part of a promising set of methods towards mitigating the environmental issues and concerns related to gadolinium contamination [180, 48, 29].

7.1.3 Extension of Modeling Capabilities and Input Data Sources

The research conducted as part of this thesis successfully demonstrates the feasibility of generating high-quality synthetic contrast-enhanced MRI (DCE-MRI) images in the realm of breast cancer using deep generative models. Despite this advancement, challenges remain, including further improving temporal coherence in multi-sequence DCE-MRI synthesis, refining contrast kinetics modeling, and validating the clinical utility of the synthetic data. Such clinical utility can further be assessed in large-scale clinical and translational studies in diverse clinical settings. To this aim, it is to be measured how generative models, such as the ones proposed as part of this thesis, can thus reliably support diagnostic workflows and/or personalized cancer treatment planning. Apart from breast imaging, this extends to further modalities where the aforementioned modeling methods are prone to result in clinically useful applications, such as, by way of example, in liver DCE-MRI [210, 221, 102]. As real-world adoption of new tools and methods are prone to follow an iterative approach, a first step towards clinical application can comprise to synthetically increase the contrast agent dose rather than generating the entire contrast synthetically. This can allow to reduce the dose of the physical injection [157]. Another feasible proof-of-concept for clinical integration is the generation of unobserved time-points of contrast-enhanced scans. For instance, based on the first DCE-MRI sequence x_{t_1} acquired after a specific time t_1 has passed since contrast agent injection at t_0 , the generative model can predict the subsequent DCE-MRI sequences $x_{t_2}, \dots, x_{t_n}$, thus potentially reducing the acquisition times. Similarly, the generative models can simulate unobserved DCE-MRI time points, e.g. $x_{t_2}, x_{t_4}, x_{t_6}$, that complement observed time points, e.g. $x_{t_1}, x_{t_3}, x_{t_5}$, enabling the

7.1. Contrast Localization and Characterization

creation of a fine-grained video showing contrast-enhancement and contrast-diffusion, rather than the common temporal extrapolation of contrast manifestations based on a few acquired time points. Furthermore, from a generative modeling perspective, multiple additional realms of further exploration exist to enhance synthetic data quality and utility. These include, by way of illustration, the joint synthesis of multiple time-conditioned DCE-MRI corresponding sequences, as well as the generation of the full 3-dimensional DCE-MRI volumes, ideally accomplishing high resolution not only in axial direction, but also in the sagittal and coronal planes. As shown, multiple conditions can be integrated into generative models, such as our proposed ContrastControlNet, enabling the addition of further complementary information into the model including clinical tabular data and/or textual descriptions of the patient, the respective medical history, biopsy information, as well as further imaging data. Such imaging data, apart from longitudinal data points, includes variations in acquisition protocols, scanner types, and field strength, but also additional imaging modalities such as diffusion weighted imaging from multiple different b-values, T2-weighted images, subtraction image sequences [191, 236], as well as non-fat-suppressed pre- and post-contrast MRI [77].

7.1.4 Measuring the Uncertainty of Synthetic Contrast

An important step towards clinical adoption is the introduction of a confidence measure to quantify the trustworthiness and certainty of a synthetically highlighted tumor. A suggested line of research is the output of a confidence map highlighting the confidence the generative model assigns to each pixel or voxel of the output image [105]. Uncertainty estimation is one methodological direction towards creating such confidence maps and can be categorized into aleatoric uncertainty (data uncertainty e.g., due to inherent data noise) and epistemic uncertainty (model uncertainty e.g. due to limited data). Uncertainty estimation methods such as Monte-Carlo Dropout [74], Deep Ensembles [120], Layer Ensembles [119], or Test-time augmentation (TTA) [216], leverage the variability between multiple predictions to extract a per-pixel uncertainty estimates. Apart from these methods, latent diffusion models such as the ContrastControlNet, contain multiple potential sources and hyperparameters that can introduce a variation of the output such as the steps of the DDIM scheduler [200] during inference, the controlnet guidance scale (i.e. the multiplier of the controlnet latent representation when added to the representation of the denoising unet [234]), or the classifier-free guidance scale (i.e. the multiplier of the text prompt latent representation [181]). Using these techniques to

vary the output of the generative model can produce useful uncertainty estimates that can lead to interpretable guidance for clinical practice. Considering model biases and best practices for responsible deployment of deep learning models in clinical practice [130], the confidence of a generative model in its prediction of how a tumor visual manifests on a medical scan can be used to aid clinical decision-making. Evaluating on a case by case basis, if the confidence is below an empirically-defined threshold, then an MRI scan with contrast agent injection can be the preferred option alongside its synthetic counterpart before deciding on the course-of-action for an individual patient.

7.1.5 Detection of Anomalous Regions-of-Interest

A further interesting avenue to explore is the comparison between real and synthetic contrast-enhanced MRI or respective subtraction images [122] to assess whether difference between the two scans can allude to a need for better understanding of the case at hand. A divergence of the synthetic scan from the real counterpart may indicate a tumor characteristic outside the training data distribution of the respective generative model or a deviation from common patient population statistics. In analogy to normative learning in anomaly detection methods [38, 18], in this scenario the non-healthy, disease-containing distribution is learned by the generative models. Hence, detecting a divergence from the predicted manifestation of a disease (e.g., wash-in and wash-out temporal patterns and hemodynamic features [61, 39], low-intensity necrotic tumor cores [2], or changes in tumor vascularization and permeability [153, 7]) can point towards a clinically relevant anomalous characteristic. Upon identification of such anomalies, these can thus be flagged for further investigation by clinical experts to explore their medical root causes and clinical impact. Apart from the characterization of a single specific lesion, unsupervised detection allows to localize multiple anomalous regions. By way of example, such methods can enhance the detection of bilateral or multifocal breast cancer manifestations as well as non-mass enhancements, as supervised lesion detection or segmentation models [166, 105] are often trained to localize only the primary lesion e.g. due to single-lesion annotation protocols of respective datasets [80, 185]. The output of supervised models can contain further bias, e.g. due to inter-observer annotation variability [105], varying annotation quality, as well as shifts in annotation protocols [6, 33]. Such biases are likely not present in the comparison of pre- to post-contrast images and respectively trained generative models, allowing the latter to be used to mitigate or adjust incoherent, imprecise, or biased outputs of the supervised models.

7.1. Contrast Localization and Characterization

7.1.6 Additional Longitudinal Temporal Pattern Prediction

The aforementioned detection of divergence from expected population statistics learned through deep generative models can further be used to assess temporal patterns. This is particularly important for monitoring and assessing treatment success in patients. Patient-specific prediction of both factual and counterfactual progression trajectories for different interventions can inform personalized treatment choices and expectations for disease evolution [59]. The response to treatments, such as Neoadjuvant Chemotherapy, is often measured based on observations on breast MRI scans that are acquired in intervals over the course of the intervention [176, 30]. Such longitudinal data acquisition is common in multiple scenarios, where patients have regular DCE-MRI examinations for purposes ranging from inclusion in screening programs to lesion monitoring, intervention planning, or disease recurrence detection. In this regard, the generative model's described in Chapter 3 and 4 can learn the additional temporal pattern over such longitudinal intervals. To date, the described models predict, based on the pre-contrast image $x_{0_{a_1}}$ acquired at timepoint $t_{0_{a_1}}$, the post-contrast images $x_{1_{a_1}}, \dots, x_{n_{a_1}}$ of the first acquisition period a_1 . However, including a temporal dimension of multiple acquisition periods $a_1, \dots, a_m$, such post-contrast images can be predicted at one or several time-points $t_{1_{a_1}}, \dots, t_{n_{a_1}}$. Such a prediction can be based on different input data points, for example, the pre-contrast image $x_{0_{a_1}}$, the real post-contrast images $x_1, \dots, x_n$, at one or several variable previous acquisition periods a , their synthetic post-contrast counterparts $\hat{x}_1, \dots, \hat{x}_n$, or a combination thereof. Such future work, predicting the additional temporal dimension of images across acquisition periods $a_2, \dots, a_m$ is envisioned to further extend the capabilities and clinical utility of the presented generative modeling methods. For instance, models can thus learn the temporal patterns of tumor growth, expected treatment response, or disease progression patterns, including time-to-event prediction such as disease occurrence [167, 172] or cancer recurrence, for which breast DCE-MRI data can hold predictive value [135, 12, 209, 40]. Similar to the methods described in Section 7.1.5, a divergence from expected temporal patterns can be indicative of clinically relevant characteristics. As an illustration, if a shrinking tumor volume is predicted by the model, but not physically observed in the real DCE-MRI data between acquisition intervals, this could describe interaction patterns between treatment and disease requiring further root cause analysis. By way of illustration, if the tumor does not respond sufficiently, this could, among multiple factors, be a potential indication of an emerging presence of a growing or specifically evolving cell subpopulation or compartments within the

tumor microenvironment (TME). Such specific cancerous cells and compartments could, due to phenotypic characteristics or mutations, acquired for example due to selective pressures caused by the applied treatment, affect the treatment response [91, 103, 228]. Due to the dynamics and heterogeneity of the TME such cell subpopulation might not have been fully sampled in biopsy probes [20]. Apart from deviations from the expected that inform clinical decision making, a further stream of application is the simulation of potential or counterfactual scenarios. At time point t_{a_1} , generative models can synthesize DCE-MRI data for time points $t_{a_2}, \dots, t_{a_n}$, thereby modeling progression patterns such as patient trajectories into the future. Such projected future scenarios can have value for clinical decision making at the present timepoint t_{a_1} . One example of an application in breast DCE-MRI is the longitudinal screening of cohorts with high-risk profile such as carriers of pathogenic BRCA1, BRCA2 mutations [116, 160]. At the present timepoint t_{a_1} , generative models are thus to learn to detect signals of nuanced precursors of breast cancer and visualize them in synthetic DCE-MRI imaging data $\hat{x}_{t_{a_k}}$ for future timepoint t_{a_k} . Furthermore, the capability of forecasting multiple potential progressions based on variation in the input data can highlight the importance of each input variable and indicate a personalized risk profile for each patient. For example, the latent diffusion model proposed in Chapter 4 retrieves a textual input containing information relevant to the modeling task at hand, such as type of contrast agent and bolus volume. Such a textual input can be readily appended in longitudinal modeling scenarios to include temporal, as well as patient (electronic health record, demographic data, genetic data, previous symptoms), disease (e.g. subtype, biopsy, BI-RADS, histology, etc) or treatment specific (type, dosage, frequency, duration, location, etc) information. When slightly varying these inputs and re-running model inference, the changes between inference outputs observed when comparing the respectively resulting synthetic images or simulations, can bear potential prognostic and diagnostic value that is to be quantified in subsequent studies.

7.2 Towards New Data and Model Sharing Paradigms

7.2.1 Solving Data Availability

The application of deep learning in medical imaging promises important improvements in diagnostic accuracy, as has been discussed in this thesis on the examples of cancer classification, unsupervised tumor localization via contrast inpainting, and breast tumor segmentation. To date, large-scale clinical deployment and adoption of such

7.2. Towards New Data and Model Sharing Paradigms

models is hindered by limitations in data availability, data privacy concerns, and the risk of patient information leakage during training and inference. Traditional data-sharing methods, which are common in other fields related to computer vision, are constrained in medical imaging due to ethical, legal, and technical barriers that are, among others, defined to protect patient privacy. Exploring solutions within the paradigm of generative model sharing as an alternative to direct dataset sharing presents a promising approach to addressing the data demands of deep learning models [206]. The utility of synthetic data, and thus of this shift, has been discussed, for instance, for large language models (LLMs), where training generative models as well as downstream task models on generated data is becoming increasingly prevalent [76, 36, 204, 218].

7.2.2 Privacy-Aware Sharing of Models and Data

While enabling collaborative research and data sharing, encapsulating knowledge from data into generative models allows medical institutions to share the essential features of a data distribution while limiting the exposure of identifiable patient information. The medigan library described in Chapter 6, presents an implementation of this concept, providing an open-source repository of ready-to-use generative models trained on a variety of medical imaging datasets. To this end, it can be argued that trained deep generative models can still leak some patient-specific information, as has been explored for instance for diffusion models [68, 32]. Furthermore, private information can be exposed via privacy attacks [136] such as ones reconstructing the images that were used as input during model training [86, 13].

Sharing Privacy-Preserving Synthetic Data Without Generative Model Weights

One option to reduce this risk while still allowing data to be shared is to distribute only the respectively generated data rather than the generative model itself. Arguably, this can further decrease privacy risks as privacy attacks often rely on access to the trained model weights [13, 86]. Indeed, based on the European Union’s General Data Privacy Regulation (GDPR) (Article 4(1) and Recital 26) [66] and related legal interpretations, data is considered non-personal if rendered synthetic to the extent that individual re-identification is no longer feasible [142]. To this end, the interpretation of “acceptable-risk” further introduces a relaxation of this definition by stating that the risk of re-identification does not need to be fully eradicated but rather reduced

to an “acceptable” level [142]. This interpretation empowers synthetic data in model training as potentially legally compliant privacy-preserving techniques, even without guaranteeing “zero” re-identification risk. To complement and demonstrate such a legal interpretation, it is to be assessed empirically to what extent the inherent properties of synthetic data, when originating from one or several different generation algorithms, provide protection against different privacy attacks. Moreover, in scenarios where synthetic data is interpreted as being shareable, it also remains an open research question to devise mechanisms that quantitatively indicate whether, when, where, and to what extent synthetic data is a viable, and potentially even preferable, alternative to real patient data.

Enhancing Utility in Scenarios with Strict Privacy Requirements

As discussed in Chapter 5, sharing of generative models or of respective synthetic data can suffice as privacy preservation mechanism in multiple use-cases. However, in particularly sensitive or regulated cases proven privacy guarantees can be a requirement, ensuring plausible deniability. Wherever such strict privacy guarantees are necessary, one option is to train generative models under a differential privacy guarantee [60] using methods such as differentially-private stochastic gradient descent (DP-SGD) [1]. Due to the post-processing theorem in DP, the synthetic data generated by a generative models trained under a DP guarantee inherently retains this guarantee. Viewing generative model training in analogy to a lossy data compression mechanism, where data is stored in model weights, future research can explore whether, and to what extent, the transition of data into generative model weights during training reduces the presence of private information. In this realm, it is an open question if the inherent randomness in generative model algorithms, including restoring [199] or sampling [83] data from random noise, can enhance the privacy of their DP counterparts to improve privacy-utility tradeoffs. First initial studies have shown the feasibility and utility of training generative models under DP-SGD based privacy guarantees [81] in the medical domain for 3D cardiac MRI synthesis [46].

Applying Local Differential Privacy After Model Training

Apart from training models under DP guarantees, a further practical alternative exists to ensure DP protection of patient data, namely, Local Differential Privacy (LDP) [109]. LDP ensures privacy by adding controlled randomness directly to data, such as medical images, rather than to model training mechanisms. LDPs potential for real-world

7.2. Towards New Data and Model Sharing Paradigms

applications in the clinical domain and beyond has been corroborated by its recent deployments on the scale of millions of users [64, 54, 15] enabling organizations to gather meaningful insights from private data. With hybrid differential privacy models being highlighted for their potential as backbone of future privacy protection frameworks [9, 15], future research can investigate privacy-utility trade-offs [47] by adjusting the strictness of LDP guarantees for individual training data points, considering factors such as data sensitivity and patient preferences from whom the data was collected. Integrating LDP with techniques such as DP-SGD presents a promising direction for enhancing privacy-utility trade-offs with personalized privacy guarantees, while ensuring a minimum global DP guarantee across the entire patient cohort in the training dataset. While LDP is primarily applied during data collection, its relevance to generative model sharing frameworks such as the medigan library lies in its potential to enhance privacy at various stages of data utilization. Specifically, where required, medigan-generated synthetic data can be subject to post-hoc differential privacy mechanisms, such as noise injection or other privacy-preserving transformations. Nonetheless, it is to be noted that post-hoc DP application can result in lower privacy-utility tradeoffs compared to models trained under DP constraints.

Sharing Differentially-Private Downstream Task Models

Apart from sharing datasets or trained generative models, a further option is to provide ready-to-use downstream task models. Despite a reduction in general applicability compared to sharing data or generative models, the dissemination of downstream task models can be less restrictive in terms of privacy concerns, while advantageous in terms of reproducibility and standardization across centers. With downstream models also potentially leaking sensitive information [86, 13], a viable mitigation measure is training such models under privacy constraints, as was previously described for DP generative model training. This is demonstrated in Chapter 5, where downstream cancer classification models are trained using DP-SGD to empirically explore the impact of different privacy guarantees and synthetic training data. The exploration demonstrated that synthetic data, used for either model pretraining or as training data augmentation mechanism, can improve the downstream task performance of the privacy-preserving classification models. This highlights the potential of synthetic data to promote broader adoption of deep learning models in clinical applications by improving privacy-utility trade-offs. This is particularly relevant in settings where (i) model performance, (ii) privacy protection, or (iii) access to diverse datasets have been critical limitations for respective development, assessment, deployment, and adoption

of deep learning models.

Towards Combining Multiple Privacy Mechanisms

In terms of privacy preservation, a further promising future direction is the combination of federated learning [150] and generative model training as shown for different medical imaging modalities by *Chang et al* [34]. Instead of transferring datasets or trained generative models, institutions can train local generative models on their own data while sharing (and receiving) only learned representations or compressed model updates. Such federated multi-center collaborative settings can further enhance privacy by training local models under DP-SGD [46], as discussed in Chapter 2. A further additional option for privacy preservation applicable in this setting is the training of deep learning models directly on encrypted data [159, 231, 44]. Such measures reduce an attackers possibility to reconstruct patient data when (a) accessing model weights, (b) retrieving respective synthetic datasets, or (c) intercepting transferred gradient values.

7.3 Trustworthy Artificial Intelligence in Healthcare: Synergizing FUTURE-AI and Generative Models

Developed in collaboration as part of this thesis, the FUTURE-AI framework is an international consensus guideline for trustworthy and deployable artificial intelligence in medical imaging [131] and healthcare [130]. For healthcare, it formulates a comprehensive set of guidelines based on consensus from 117 interdisciplinary experts from 50 different countries, including AI scientists, clinical researchers, and social scientists. The set of guidelines are structured into seven guiding principles including a total of 30 recommendations. The principles, apart from a *General* Principle (7 recommendations), include the **F**airness (3), **U**niversality (4), **T**raceability (6), **U**sability (5), **R**obustness (3), and **E**xplainability (2) principles.

7.3.1 General: Privacy via De-Identifying Data Synthesis

Applicable to generative model solutions, recommendation 2 of the General FUTURE-AI Principle guides to *implement measures for data privacy and security*. As discussed in the previous section 7.2.2, using data generated by generative models rather than real data addresses data privacy requirements. This is aligned with General recommendation 5 suggesting that one should *identify and comply with applicable AI regulatory*

7.3. Trustworthy Artificial Intelligence in Healthcare: Synergizing FUTURE-AI and Generative Models

requirements during AI development and deployments. Such regulatory requirements, stipulated for example by the GDPR [66] or HIPAA [212] regulations, can require privacy protection of personally identifiable information. As investigated in Chapter 5 and reviewed in Section 7.2.2, generative models can be a useful method to de-identify a data source by providing and sharing only its synthetic data and/or by training under strict differential privacy guarantees.

7.3.2 Fairness: Via Balanced Dataset Generation

In terms of the Fairness principle, FUTURE-AI recommends that one *defines any potential sources of bias from an early stage* (1), *collect information on individuals' and data attributes* (2), and *evaluates potential biases and, when needed, bias correction measures* (3). For instance, a downstream task model trained predominantly on data from a single ethnic, gender, or demographic group is prone to contain limitations in generalizing to other underrepresented populations. This can be particularly concerning in oncology, where tumor characteristics and imaging biomarkers can vary significantly across genetic and demographic groups [4, 78, 33, 107]. If diagnostic models are biased towards specific populations or types of cases from the training distributions, they can demonstrate lower performance levels in other subpopulations or types of cases. For example, a collaboration supervised by the author of the present thesis demonstrated that fairness across populations with different skin colors is questionable and likely not correctly assessed in current skin lesion classification models in dermatology [107]. A further collaboration by the author showed that the performance of models detecting breast masses in screening mammography varies for patient populations depending on their breast density [79]. After defining, quantifying, evaluating and reporting such biases, one potential next step is the mitigation of these biases, where one solution is to generate synthetic data representing patient groups that are otherwise underrepresented in the respective dataset. As was investigated in collaborative works [6, 107, 198, 79] conducted alongside this thesis, adding such synthetic data to the training datasets of supervised deep learning models can enhance performance and fairness across patient subgroups. This emphasizes the potential of open sharing of generative model weights, as discussed in Chapter 6, which can foster broad and equitable access to synthetic datasets across institutions including those in resource-constraint settings. Such access to the capabilities of pretrained generative, as well as downstream task models, allows research institutions and hospitals to balance their respective datasets and to analyze and ensure that their models benefit all sub-groups of their patient population.

7.3.3 Robustness: Enhancement via Synthetic Data

Recommendation 3 of the Robustness principle suggests to *evaluate and optimize robustness against real world variations* for both the research stage as well as the deployment stage of AI solutions in healthcare. As elaborated in Chapter 2, generative models are capable of simulating variations of respective data using techniques such as image editing, conditional synthesis, and image-to-image translation. This enables the generation of data that bridges existing domain gaps, populating low-density parts of the data distribution at hand. One illustration of this principle includes the translation of T1-weighted pre-contrast breast MRI to respective dynamic contrast enhanced (DCE)-MRI using generative adversarial networks in Chapter 3. For clinical scenarios where a tumor segmentation model encounters discrepancies between the training and inference data distributions, i.e. training on pre-contrast images while having to perform inference on post-contrast images as explored in Chapter 3, the segmentation performance can be adversely affected due to such domain gaps. However, incorporating synthetic post-contrast MRI volumes into the pre-contrast training dataset can notably improve segmentation performance, highlighting both the existence and impact of the domain shift, while also corroborating the potential of synthetic data in mitigating such discrepancies.

7.3.4 Explainability: Via Controlled Variations in Input Conditions

FUTURE-AI also recognizes the need for intuitive explanations allowing researchers, clinicians, and patients to interpret and scrutinize the rationale behind a model's predictions. Taking this under consideration, the second recommendation of the Explainability Principle suggests to *evaluate explainability with end users, for example, on correctness and impact on users*, after *defining the need and requirements for explainability with end users* respectively in the first explainability recommendation. In this regard, a visually interpretable approach to explaining model predictions involves systematically varying the model's input and observing the corresponding changes in its output. The generation of counterfactual data points can thus enable controlled input variations, allowing for the exploration of alternative scenarios beyond the original case. By conditioning the generative process on specific variables, such as patient age, health status, or disease progression markers, the underlying factors influencing model decisions can be analyzed. For instance, generative models can synthetically alter disease manifestations or anatomical structures in organs such as the brain [168] or heart

7.4. Trustworthy Artificial Intelligence in Healthcare: Synergizing FUTURE-AI and Generative Models

[31] based on conditioning variables like patient age. Feeding these altered images into a secondary downstream model enables the assessment of prediction changes, thereby enhancing model interpretability and supporting personalized clinical decision-making. As generative models gain traction in clinical applications, their role as explainability tools through counterfactual data generation can also improve usability and, overall, trustworthiness of AI solutions in healthcare.

7.3.5 Explainability: Via Counterfactual Synthesis

Recommendation 3 of the Usability Principle encourages to *provide training materials and activities such as tutorials and hands-on sessions*, facilitating know-how and ease of adoption for AI tools in healthcare. In this realm, generated counterfactual data can be useful as education material for clinical users of AI tools to interactively and visually explore a model's data sources and limitations. Even beyond tool usage, counterfactual samples can also enhance the training curriculum of health professionals, exploring simulated low-probability patient scenarios and augmenting the number of cases and variations available as education material.

7.3.6 General: Definition of Adequate Evaluation Plans

Generative models are to become increasingly integrated into clinical practice, for instance, as direct problem solvers for downstream tasks such as virtual contrast synthesis discussed in Chapters 3 and 4, as tools for patient cohort anonymisation and data sharing as suggested in Chapters 5 and 6, or as methods for augmenting the training data of other downstream task models elaborated in Chapters 3, 5, and 6. This highlights the growing importance of research into meaningful, reliable, and systematic evaluation of both generative models and of the data they produce. To this end, FUTURE-AI envisions in its fourth recommendation within its General Principle, the importance of *defining adequate evaluation plans, e.g., datasets, metrics, and reference methods*. However, research into the design of such concrete evaluation plans has been limited, in particular for generative models and for their evaluation as they are adopted in clinical settings. This motivates the exploration of evaluation frameworks, dimensions, and metrics as part of the present thesis, with respective results being summarized and discussed in the subsequent Section 7.4.

7.4 The Intricacies of Evaluating Generative Models

Apart from proposing modeling methodologies and applications, this thesis places further emphasis on the evaluation of generative models and corresponding synthetic data. Ensuring the quality, reliability, and clinical applicability of synthetic data is critical for its integration into medical imaging workflows. With proposed evaluation methods distributed across Chapters 2-6, this section constitutes a consolidated discussion of these methods, providing a structured analysis of generative model assessment through the defined frameworks, criteria, metrics, and domain-specific evaluations.

7.4.1 Evaluating Synthesis Studies using SynTRUST

Table 7.1: Summary of the SynTRUST framework defined in Chapter 2 alongside its dimensions and building blocks for evaluating the trustworthiness of studies proposing synthetic data or respective data generation methods in medical imaging.

Dimension	Description and Rationale	Concrete Measures
Thoroughness	Ensuring rigorous and well-validated experiments, with generalizable results and comparability across models, based upon which generally applicable conclusions can be drawn.	Minimum test set size, multi-metric reporting, multiple result validation runs, fair baseline comparison, principled benchmark definition, significance testing, ablation of key components of a method, effect of varying training set size.
Reproducibility	Ensures that studies can be repeated and verified by independent researchers, promoting transparency in reporting methods, datasets, and models	Detailed reporting of design decisions, public availability of dataset, public availability of software code, public availability of model weights.
Usefulness	Practical applicability of synthetic data and models, approximating how well synthetic data serves real-world use-cases	Synthesis method usability testing, quantitative quality measurement, qualitative quality measurement, mode collapse analysis.
Scalability	Transferability across datasets, institutions, and clinical applications, measuring generalization across a-priori known and unknown variations in real-world settings	Real-world representing data, multi-dataset evaluation, multi-centre evaluation, multi-downstream task evaluation, downstream task robustness evaluation.
Tenability	Trustworthiness, fairness, and privacy considerations, addressing concerns regarding potential biases and risks inherent in or associated with models and their outputs.	Condition adherence testing, bias awareness analysis, model hallucination tendency analysis, fairness variation testing, privacy preservation testing.

7.4. The Intricacies of Evaluating Generative Models

Considering existing frameworks for reporting healthcare-related AI studies such as FUTURE-AI [129], TRIPOD+AI [43], and TRIPOD-LLM [75] focusing on general-purpose tools, supervised learning methods, and large language models, this thesis identified a gap in validation and reporting practices specific for image synthesis. Addressing this gap, the SynTRUST framework, introduced in Chapter 2, provides a comprehensive approach for evaluating studies that utilize or propose data synthesis. Consisting of 26 concrete measures distributed across its 5 core principles, the SynTRUST framework is summarized and discussed in Table 7.1 with its core principles being **T**horoughness, **R**eproducibility, **U**sefulness, **S**calability, and **T**enability. As visualized in Chapter 2, the number and variety of publications in the field of deep generative models in computer vision and medical imaging has been notably increasing. This demonstrates the timeliness and importance to define frameworks for a specific and principled analysis of such types of studies. This further aligns with the emerging need and ultimate objective of demonstrating the technology maturity and readiness for clinical adoption of such models. To conduct clinical assessments and trails of generative model, the validation-specific measures defined by SynTRUST can be consulted as a solid indicator for deployment maturity. However, further dimensions exists and are to be taken into consideration going beyond the technical validation of the models. These include all areas where deployed models constitute a deviation from the prevailing status quo, e.g., impacting clinical workflows, patient assessments, tool usage and automation, as well as having societal or environmental effects.

Risks of Automation Bias

For instance, one particular real-world impact of clinical deep learning models is automation bias, which occurs when clinicians over-rely on AI recommendations and potentially overlook their own judgment. Among other studies [178, 82], *Dratsch et al.* [58] recently showed that radiologists interpreting mammograms were significantly influenced by incorrect AI suggestions. For instance, the diagnostic accuracy of the group of less experienced radiologists notably plummeted from 79.7% to 19.8%. Thinking of automation bias in terms of generative models, one hypothesis is that frequent visualization and assessment of synthetic images, e.g. to simulate patient trajectories or virtual contrast agents, can reduce the awareness of clinicians to detect nuanced synthetic anatomical deviations from real data. Frequent observation of image-to-image translation outputs, for instance, could result in decreased capabilities of distinguishing between correct and hallucinated image contents. Paired with disproportionate trust in such outputs, such an automation biases can negatively impact clinical decision-

making and highlight the need for training and safeguards to mitigate automation bias. Ensuring that AI models serve as a clinical support tool rather than a replacement for clinical expertise is, thus, a critical dimension for evaluation of models, calling for adjustments in study designs and corresponding assessment measures and metrics.

Increasing Efficiency and its Causal Mechanisms

Other evaluation dimensions are to consider cost-effectiveness and efficiency of generative AI solutions, particularly when integrated into clinical workflows. For example, measuring a reduction in the time it takes to conduct a clinical task by using a deep generative model can potentially allude to increases in efficiency and cost effectiveness [110]. However, it is to be scrutinized whether a reduction in time for task completion is rather caused in part, or entirely, by other factors associated to cost-cutting and automation initiatives that can result in a reduced quality of care. Despite the potential benefits of deep learning models, reducing human autonomy or the level of human supervision of AI tools and of clinical decision-making, can become a risk for a declining quality of the related diagnostic tasks, clinical interventions, and patient interactions. To this end, such risks and effects are to be analyzed and evaluated, facilitating an awareness whether a need emerges to define mitigation strategies or modifications of AI-assisted clinical workflows to ensure improved health outcomes. For instance, an earlier detection of breast cancer not only improves health outcomes, but also leads to a substantial reduction in respective treatment costs [22].

Sustainability of Model Training and Inference

Further normative assessments and trade-offs exists that require in-depth evaluation of deep learning models in general and of deep generative models in particular, such as their resource efficiency and environmental impact. To this end, recent large-scale deployments of autoregressive large language models (LLMs) have drawn attention to the vast energy consumption of such generative models [187, 96]. If the availability and size of resource-intensive general-purpose models continues to expand, further encompassing multi-modal generative models [205], vision-language models [25], and agentic systems [196] alongside LLMs, overall energy consumption levels can be expected to continue to rise substantially. In this regard, resource efficiency and sustainability, for instance in terms of carbon emission [192] and water consumption [134], of deep generative model training and inference is a further dimension of impact that can be integrated into model evaluation frameworks. Once a model's environmental footprint

7.4. The Intricacies of Evaluating Generative Models

is quantifiable, comparable, and embedded within standard evaluation practices, it becomes easier to assess trade-offs associated with model training and adoption. This, in turn, enables the development of strategies to optimize these trade-offs. For instance, the *CodeCarbon* tool [45] provides estimations of CO₂ equivalents produced by executing compute-intensive code, supporting sustainability-aware decision-making in model development. To this end, as part of a collaborative effort associated with this thesis [70], *CodeCarbon* was used to analyze the resource efficiency of deep learning models. This allowed to demonstrate that comparable performance levels on clinical downstream tasks can be achieved while reducing CO₂ equivalents and training runtime by approximately 50% through the use of progressively increasing input patch sizes within a curriculum learning framework [70].

Transparency in AI Supply Chains

Another additional evaluation dimension comprises the ethics of generative models, extending beyond technical performance and including, for example, labor conditions within AI supply chains [8]. Studies on outsourcing of data annotation processes describe multiple issues within the supply chains of annotating the data of generative models, including exploitative labor practices, low-wages, and job insecurity, challenging the ethical integrity of AI development initiatives [155, 156]. To ensure responsible AI deployment in medicine and beyond, evaluation frameworks are to assess the supply chains leading to the provision of AI solutions, fostering fair sourcing and working conditions, as well as transparency, accountability, and social responsibility across AI value chains.

Overall, the SynTRUST framework represents a foundational step toward the comprehensive evaluation of generative models and their publications while identifying key directions for future research. Alongside the previously discussed aspects, these directions include ethical, social, and clinical impact metrics, traceability measures e.g. for assessing synthetic data provenance and modifications, as well as regulatory evaluations informing stakeholders about legal, clinical, and societal implications. Patients, for example, represent one such key stakeholder group that requires a (a) rigorous evaluation of AI solutions, (b) transparency in their application, and (c) the involvement of a radiologist in order to establish trust in AI, as was recently analyzed for the case of breast cancer screening [104].

Table 7.2: Summary, applications, and challenges of evaluation dimensions for synthetic data in medical imaging based on Chapter 3.

Evaluation Criterion	Description	Rationale	Example Applications	Challenges
Fidelity	Measures how well synthetic images replicate real-world imaging characteristics, including visual realism, texture, shape, and statistical properties.	Ensures synthetic images preserve desired target distribution features required for medical diagnosis and treatment planning.	Generating the exact features expected in real data in paired image-to-image translation	Achieving high realism while simultaneously avoiding overfitting or memorization of a specific training subset of the real data distribution.
Diversity	Assesses variability in generated images, ensuring representation of anatomical, pathological, and contextual variations.	Implicit measure of the density of the model distribution and potential mode collapse, to ensure synthetic datasets capture the full spectrum of variability found in respective real images.	Generating synthetic data across different disease stages, patient demographics, and imaging conditions.	Maintaining diversity without introducing unrealistic variations or artifacts.
Condition Adherence	Evaluates whether synthetic images accurately reflect predefined attributes, constraints, or conditions specified during generation.	Ensures generated data aligns with intended attributes such as tumor presence, lesion size, or imaging modality characteristics.	Conditional GANs and diffusion models generating contrast-enhanced images based on conditional inputs such as pre-contrast images or labeled lesion categories.	Maintaining precise adherence to input conditions while avoiding confounding factors or hallucinated artifacts, and overly focusing on conditions causing mode collapse.
Utility	Measures how well synthetic data supports downstream applications, including as augmentation of machine learning model training data	Ensures that synthetic images contribute effectively to specific practical tasks or to models solving those tasks.	Enhancing tumor classification/segmentation performance when trained on synthetic data or a combination of synthetic and real data.	Ensuring synthetic images are useful for (multiple) tasks, despite training the synthetic data generator on the generic objective of approximating or mapping to the real-world data distribution.
Privacy	Assesses whether synthetic images protect sensitive patient information while maintaining clinical relevance.	Prevents patient re-identification while allowing secure AI model training and dataset sharing.	Make (synthetic) datasets shareable e.g. by applying differential privacy to model training, protecting them against privacy attacks (e.g. membership inference attacks).	Balancing the tradeoff between privacy with image fidelity and utility, ensuring synthetic data remains useful.
Fairness	Ensures synthetic data represents diverse patient populations, avoiding biases that could lead to inequities.	Prevents biased AI decision-making and ensures equitable performance across demographic and clinical subgroups.	Generating balanced datasets with equal representation of different patient demographics, disease types, or imaging protocols.	Detecting and mitigating hidden biases in training data that propagate into synthetic images.

7.4.2 Quality Assessment for Medical Image Synthesis

Shifting from a holistic evaluation of synthesis studies to an assessment of synthetic data itself, the present thesis identifies multiple desirable properties and characteristics of synthetic data. Based on the definition and discussion in Chapter 3, such characteristics of medical synthetic data can be structured into six evaluation dimensions,

7.4. The Intricacies of Evaluating Generative Models

namely, (i) Fidelity, (ii) Diversity, (iii) Condition Adherence, (iv) Utility, (v) Privacy, and (vi) Fairness. Table 7.2 summarizes and discusses these dimensions alongside their applications and respective challenges. Overall, these dimensions help to comprehensively determine whether synthetic data accurately reflects real-world distributions and applications while minimizing biases and ensuring regulatory compliance. Providing a structured framework of evaluation dimensions of synthetic data is to increase trust in AI-generated datasets, particularly when considering applications in high-stakes environments such as healthcare and biomedical research.

Unexplored Interdependencies Between Dimensions

Nonetheless, some challenges remain in the assessment of these dimensions. Firstly, it is to be noted that these six dimensions are often interrelated and often positively correlated. For instance, intuitively, if a synthetic dataset exhibits a high diversity, this will likely also cause a high score in fairness. Similarly, one may readily assume that a higher fidelity of the synthetic data will also result in higher utility. In terms of an increase of synthetic data privacy, on the other hand, it is likely that this causes a decrease in utility, as recent studies explored [81, 46]. Despite these seemingly intuitive assumptions, the cross-correlations between evaluation dimension are a-priori unknown for each model and dataset. Scores per dimension can even be mutually exclusive exhibiting a trade-off, as was recently shown for image fidelity, variety, and privacy [230]. On the other hand, it was further possible to produce images that simultaneously exhibited high utility and privacy, as well as high utility and low fidelity [230]. In general, the relationship between the score of one dimension and that of another remains unknown. This uncertainty further suggests that trade-offs between dimensions may not be static but could vary dynamically as one score changes. To measure such scores, commonly only approximations are observable, as synthetic data evaluation metrics serve as surrogate measures for specific evaluation dimensions. For example, while privacy is often assessed based on a differential privacy guarantee [60] or approximated by results of different privacy attacks, utility is typically estimated using downstream task performance. To this end, utility can vary based on the multitude of possible metrics and tasks, as well as the choice and hyperparameters of a second downstream task model trained with the synthetic dataset. Moreover, the Fréchet Inception Distance (FID) [89] is often used as proxy for both fidelity and diversity. A recent study found that different variations of the FID metric produced substantial inconsistencies in their correlation with radiologists' assessments in a visual Turing test of synthetic datasets [224]. This finding further underscores the challenge of accurately

measuring evaluation dimensions in synthetic data assessment.

Loss Function Designs for Multi-Dimensional Objectives

With future work aiming to increase the understanding of the interconnectedness and functions between evaluation dimensions and their corresponding evaluation metrics, another research question is how the objectives posed by each dimension can be ideally integrated into model training. By way of illustration, this can include the ideation, definition and integration of dedicated loss functions each optimizing its respective evaluation dimension. This thought further leads to the subsequent research question to define how to weight these loss functions in comparison to each other. Apart from the challenge of finding technical best practices on how to achieve an overall best possible tradeoff between all of these training objectives, the weighting of such objectives inherently involves a normative component, initiating a philosophical debate. This is exemplified by the challenge of determining, for instance, whether greater importance should be assigned to privacy over fairness, fairness over utility, or utility over fidelity.

7.4.3 Metric Contributions for Image Synthesis Assessment

Further shifting from the evaluation dimensions of synthetic data to concrete measures, this thesis proposes several novel metrics to assess generative models and their respectively generated dataset. These metrics, introduced in Chapters 3, 4, and 6, are summarized and discussed in Table 7.3.

Noting limitations in current synthetic data evaluation metrics, one question is how to ensure that metrics capture the nuances of medical data. For instance, FID [89] is based on a feature extractor network [207] pretrained exclusively on natural images [50]. In this regard, this thesis proposes domain-specific metrics such as the FID_{Rad} in Chapter 6, which computes the Fréchet Distance based on a feature extractor network trained on a varied set of radiology images [151]. A further novel metric introduced and validated as part of the present thesis in Chapter 4 is the Fréchet Radiomics Distance (FRD), which calculates the Fréchet Distance based on extracted [213] and normalized radiomics features. Beyond being domain-specific, key advantages of the FRD include the interpretability of its handcrafted features, its flexibility in being extractable from both two-dimensional and three-dimensional medical images, and the demonstrated utility [121, 30, 145, 20, 94, 202] of radiomics features as versatile imaging biomarkers across a wide range of pathological conditions, health outcomes, and patient characteristics. In this regard, evaluating images based

7.4. The Intricacies of Evaluating Generative Models

Table 7.3: Overview of generative model evaluation metrics in medical imaging proposed in the present thesis across chapters 3, 4, and 6.

Metric	Description	Rationale	Limitations
FID_{Rad} (Chapter 6)	A variation of the Fréchet Inception Distance (FID) [89] that uses a medical imaging-specific feature extractor pretrained on RadImageNet [151], instead of an ImageNet-trained [50] InceptionV3 [207].	Focus on the domain relevance of FID, ensuring a domain-specific distribution similarity assessment based on features learned from medical data.	Two-dimensional images required, without applying further optimization, FID _{Rad} has limited correlation with human observer assessment in current implementation [224].
FID Ratio (Chapter 6)	A ratio of FID comparing the naturally occurring variation within sets of real data to the variation between sets of real and synthetic data	Using relative rather than absolute FID values reduces sensitivity of FID by canceling out external factors such as dataset variation, subset size, applied normalizations allow for better comparability of FID between e.g. applications, organs, modalities, and datasets.	The correlation with image quality and task utility is to be further analyzed to establish common ranges and intuition for interpretation of calculated FID ratios.
SAMe Score (Chapter 3)	Scaled Aggregate Measure (SAMe) integrates multiple generative evaluation metrics into a single, weighted score to assess synthetic data quality holistically.	Perceptual, semantic, and pixel-wise metrics on image and distribution levels can show contrary results, but are complementary meaningful motivating their comprehensive assessment as an aggregate score.	While particularly useful for checkpoint selection during training, weighting of different components is task-dependent and requires careful calibration, which may limit its generalizability across modalities.
FRD (Chapter 4)	The Fréchet Radiomics Distance is a medical imaging-specific metric that extends FID using radiomics features instead of deep learning embeddings	Providing a domain-specific quality assessment by measuring distributional similarity in radiomics feature space widely used in clinical analysis, FRD features correlate with image quality, are domain-specific, interpretable, work for both 2D and 3D datasets, and include well-established clinically meaningful imaging biomarkers	Further investigation needed for assessing sensitivity to noise and preprocessing variations to ensure reproducibility.

on imaging biomarker distributions constitutes a technique that merges the concepts of clinical utility, image quality, and image diversity, as demonstrated in subsequent work expanding the experimental validation of FRD [113].

Automated Assessment of Domain Shifts

Beyond synthetic data evaluation, the comparison of imaging biomarker distributions has broader applications, highlighting the potential of further validation of the FRD in these diverse contexts. This includes the automated assessment of domain shift between datasets, ensuring the robustness of imaging studies in general. One respective example is the evaluation of imaging protocols by comparing distributions from repeated scans under identical conditions to validate their reproducibility. On the other hand, FRD can be useful to analyze cross-center standardization, quantifying the extent to which

images acquired from different hospitals or research centers remain consistent despite variations in equipment or scanning protocols. This can enhance the awareness of potential biases introduced by inter-center differences that ultimately can impact the generalizability of image-based models and clinical decision-making.

Monitoring Consistency Across Processed Datasets

Similarly, comparing imaging biomarker distributions allows to monitor image preprocessing pipelines, for instance by assessing the consistency across normalized, rescaled, resized, denoised, or cropped datasets. In this regard, FRD can detect image degradations as shown in Chapter 4 by comparing distributions before and after processing steps to identify artifacts and changes in image quality. Another application is the validation of super-resolution and image restoration techniques, examining whether reconstructed images maintain the biomarker distributions of real high-resolution images. Similarly, data compression techniques can be evaluated to determine whether compressed datasets retain the statistical properties of their original counterparts. Beyond evaluation, a further research question involves whether image modification techniques can be dynamically optimized using measures such as FRD, including investigations into adjusting such techniques based on changes in biomarker distributions. Such changes can be measured and optimized, for example, via the feature space distance between a given sample and a target biomarker distribution or between specific reference points within that distribution.

Evaluation of Annotation Variability

Since FRD allows passing segmentation masks alongside images to the radiomics feature extractor, additional applications emerge in the realm of segmentation evaluation by comparing features of the same images derived from different annotation masks. One such application is to ensure consistency in organ or tumor annotations by comparing feature distributions across different annotations from radiologists. By way of example, this can enable the quantification of the reproducibility of segmentation tasks, as well as intra-observer and inter-observer variability. Apart from manual annotations, this approach is further applicable to automated segmentation algorithms, examining their alignment with expert ground-truth annotations. Current standard metrics for evaluating segmentation and object detection, such as the Dice coefficient or the Intersection over Union (IoU), quantify geometric overlap between selected regions but do not capture the semantic or downstream impact of those differences. In contrast,

7.4. The Intricacies of Evaluating Generative Models

FRD provides a means to assess how variations in annotations influence feature representations and potentially affect downstream tasks. For example, a segmentation mask that includes slightly more surrounding healthy tissue (i.e., is zoomed out) can influence radiomic features and respective downstream model predictions more than one that is slightly too focused (i.e., zoomed in), as the latter captures only tumor tissue with relevant intensity and texture-related features. Thus, FRD can offer a more task-relevant perspective on segmentation quality compared to traditional spatial overlap metrics. With this in mind, FRD can further be used to evaluate and track segmentation and object detection outcomes after and throughout model training, as well as in curriculum learning scenarios. Detecting model improvements and deviations over time, FRD can be used both as a standalone measure as well as a complement to traditional pixel-wise metrics, such as the Dice coefficient, or set-based metrics, such as the Hausdorff distance.

Assessment of Test-Retest Variability

It is to be further examined to what extent FRD can detect consistency of imaging patterns and respective shifts, such as prevalence and acquisition shifts [33], over time by comparing distributions from repeated scans of the same patients in test-retest setups [214, 174]. Similarly, this approach has the potential to detect disease-specific distribution shifts, such as changes in disease manifestations [33], identifying alterations in image distributions associated with medical conditions. This extends to monitoring disease progression, where analyzing changes in biomarkers can enable tracking of disease evolution and treatment responses. Comparing such distributions across patient groups (e.g., age, sex, ethnicity) can help identify biases and population shifts [33], where variations in biomarker manifestations can potentially be linked to medically relevant findings and implications.

Detection of Adversarial Attacks

Metrics for imaging biomarker comparisons on image and distribution-level can further emerge as technique to detect adversarial examples. Adversarial attacks manipulate images to mislead machine learning models while maintaining a visually imperceptible appearance [84, 5, 177]. In this regard, a reasonable hypothesis is that imaging biomarkers can be used to identify subtle deviations in the feature space that do not become readily apparent in semantic, perceptual, or averaged pixel-wise comparisons. By comparing the feature distribution of seemingly identical images or datasets, imag-

ing biomarkers can thus be useful to detect discrepancies introduced by adversarial perturbations [14]. Future work is to examine the respective sensitivity of biomarker-based distribution comparisons to adversarial perturbations to gain insights into their effectiveness as a defense mechanism against adversarial attacks.

7.4.4 Emerging Generative Model Evaluation Paradigms

Widely-acclaimed demonstrations and recent advances of the capabilities of deep-learning [41, 115, 152, 83, 194, 195, 215, 52, 173, 106, 182, 175] have sparked a Cambrian explosion of AI development, with global research endeavors rapidly advancing the state-of-the-art of deep learning in general, and generative models in particular. The respectively increasing complexity and capabilities of emerging model paradigms highlight the need for further advancing evaluation mechanisms as well, in order to assess such models and ensure alignment with their intended objectives.

Evaluation of Multi-Modal Generative Capabilities

Extending to the field of medical imaging, generative models are becoming increasingly multi-modal as has been illustrated on the example of joint text and image model inputs in Chapter 4. Multi-modal input and output mechanisms allow for novel interaction mechanisms between users and AI models in medicine. Examples include haptic, voice and, in particular, chat interfaces for models that are capable of producing outputs based on nuanced patterns in input prompts [233, 132, 139, 133]. This ongoing paradigm shift towards multi-modal model interaction requires ideating new avenues in terms of the evaluation of such models and their generated outputs. For instance, it is an open question whether the joint synthesis of multiple output modalities rather requires a per-modality evaluation or a joint evaluation of all modalities of the output. In the latter case, it is to be defined how these multiple modalities are to be assessed together considering that their desired level of correspondence to each other and to the outputs can vary between models, applications, and use-cases. This is paired with an increase in complexity of modalities, where models can, for instance, output entire medical histories and patient records in textual formats [132] as compared to binary labels to evaluate an input image. When multiple input modalities are considered simultaneously to evaluate the quality and adherence of output images to given conditions, determining an expected appearance of such outputs can become unclear, subjective, and inconsistent. In addition, since often multiple factually correct mappings exist from the same inputs to different outputs, assessing and comparing

7.4. The Intricacies of Evaluating Generative Models

these outputs objectively becomes increasingly intricate. Among others, ambiguity of the correctness of outputs is due to the complexity to define expectations for reasonable input condition co-correlations, a model’s respective multi-modal pattern recognition ability, as well as desired model-internal weighting mechanisms for each of its inputs.

Defining robust evaluation mechanisms becomes progressively more demanding in the presence of ambiguity arising from multi-modal model inputs. This is depicted by the hypothetical scenario where a model input consists of both a patient’s histopathology scan and a corresponding radiology scan. If, for instance, a pathology image exhibits patterns indicative of a benign finding, while features from the radiology scan suggest malignancy, the discrepancy could stem from a biopsy sample that is too small or outdated, or it may indicate an error in the radiology assessment. Without additional guidance, it can remain uncertain to the model which of these modalities to prioritize during processing. Further, it is challenging to determine how the model’s corresponding output (e.g., a generated image, annotation, or diagnostic label) is to be evaluated given expressive and potentially inconsistent inputs, where, moreover, such inconsistencies may well be unknown to the person using the model or assessing its predictions. Similar challenges can arise in multiple settings where nuanced input modalities can simultaneously express multiple related concepts. For instance, considering textual inconsistencies, gaps, and variations in input radiology reports, it is unclear how generative models tasked with producing corresponding radiology scans are to resolve such inconsistencies. These complexities allude to the need for evaluation frameworks that account for uncertainty, potential diagnostic errors, and the complexity, potential contradictions and inherent limitations of pattern recognition across multiple merged modalities.

Taking the aforementioned aspects under consideration, it becomes apparent that further work is needed on evaluation mechanisms that go beyond current distribution-level [89] or image-level [219, 235] comparison methods. The latter are limited, for instance, in terms of measuring the adherence to multiple and multi-modal conditions, as well as particularly in cases where the generation is based on highly nuanced and expressive conditional input modalities such as long free-text formats.

Evaluation of Sampling in Sparse Regions of the Distribution

A wide range of deep generative models exists including variational autoencoders (VAEs) [111], generative adversarial networks (GANs) [84], diffusion models [199, 201, 92] and their variants in latent space [181], flow matching models [137], normalizing flows [179, 55, 56], as well as auto-regressive transformer models [215]. One common

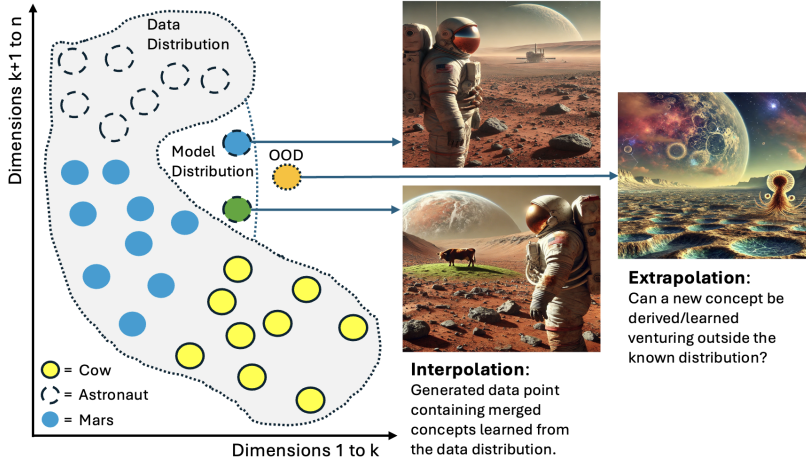


Figure 7.1: Creativity in deep generative models: Illustrative visual depiction of interpolation and extrapolation capabilities of generative models. Out-of-distribution (OOD) samples containing concepts that are not readily present in the training data distribution, but are extrapolated abstractions thereof. The demonstrated images were generated by a DALL·E model using the ChatGPT 4o interface (version date: 12.11.2024).

characteristic of these generative modeling methods is that they either implicitly or explicitly learn a data distribution. This allows to sample from the learned model distribution that approximates the real data distributions, resulting in the ability to produce realistic samples from that distribution. This process is conceptually analogous to learning the probability density function (PDF) of the target distribution, particularly in models where the density can be explicitly estimated. The PDF characterizes the likelihood of different data points occurring within the data space. One particularly useful feature of generative models can be their ability to learn to capture and draw samples from low-density regions of the PDF, i.e. areas that are underrepresented or sparse in the real dataset. With future work advancing the possibilities of quantifying and modeling these regions effectively, generative models can enhance the coverage of the target distribution and improve the generalization of downstream models trained on both real and synthetic data.

Emerging Interpolation Capabilities of Generative Models

Overall, future work into the ability and functionality within generative models to (a) capture low-density regions, (b) interpolate between data points within a distribution, (c) merge multiple concepts, and (d) exhibit emergent capabilities are important lines

7.4. The Intricacies of Evaluating Generative Models

of research due to their wide-ranging implications in healthcare and beyond, as, in part, discussed in Chapter 2. For example, in the context of emerging capabilities, generative models, such as large language models, can acquire general-purpose skills that give rise to novel behaviors that reflect a form of machine creativity and can enable new applications such as complex reasoning and data manipulations [220]. In parallel to advancing modeling techniques, a critical next step is to also include these learned behaviors in evaluation frameworks, either by devising methods to directly measure them or by defining suitable surrogate metrics that allow to effectively capture and analyze respective aspects.

In this regard, Fig. 7.1 presents two generated examples illustrating the interpolation capabilities of a general-purpose open-access vision-language model merging visual concepts such as *Planet Mars* and *Astronaut*. The prompt for the top image requested *an image of an astronaut after landing on planet Mars, having arrived in a rocket, with Earth and the universe in the background*. This was further refined to request *a more realistic image as if actually taken on Mars with a normal camera*, resulting in the final output. The bottom image in Fig. 7.1 was generated using the prompt *now generate the same image but with a realistic cow eating grass growing on one particular part of the surface of Mars in the background*. This prompt combines multiple semantically distant concepts such as *Cow*, *Planet Mars*, *Grass*, *Astronaut*, *Surface*, which can be observed to be effectively integrated to produce a coherent and visually realistic, yet fictional, image. As there are, to date, no documented instances of manned expeditions to Mars, the interpolation demonstrated by the generative model can be interpreted as a form of creative synthesis, producing novel visual content. However, these newly merged concepts are not entirely novel, but rather grounded in information the model has already encountered in its training data, albeit as separate, uncombined concepts.

Potential Extrapolation Capabilities of Generative Models

In terms of generating entirely novel content, it can be hypothesized that generative models, in a process arguably to some extent analogous to human cognition, can become capable of extrapolating beyond the concepts represented in their training data. This goes beyond the discussed model capability of filling low-density regions of the data distribution, enabling near-complete coverage and high-quality interpolation between data points. Rather, extrapolation, as a more speculative yet transformative capability, involves a model’s ability to generate data points that fall outside the boundaries of its training distribution. Fig. 7.1 presents an example of a simple prompt-based attempt to generate such an out-of-distribution data point using a recent

general-purpose vision language model. The prompt was to *generate an image of an unseen planet in another universe with an unknown creature that follows an unknown process for something along the lines of preservation and life support. The image needs to be realistic as taken with a camera and cannot contain any known elements, materials, concepts or life forms.* While the intention was to fully exclude familiar concepts, the resulting image still contains recognizable structures such as clouds, mountains, and surface textures, highlighting the inherent challenge in both defining as well as detecting true out-of-distribution extrapolation capability in generative models. Future work on efficient training mechanisms towards extrapolation is to represent substantial advancements, addressing key limitations of current generative models in particular and artificial intelligence in general. This is due to the multiple implications of extrapolative abilities, which, among others, can become a foundational pillar for reasoning, artificial creativity [23, 184, 217], and, potentially, guide the development of ideas towards artificial curiosity [190]. Abilities such as generating new knowledge, formulating hypotheses, and proving existing or novel theorems represent a cornerstone of scientific advancement and hold the potential to push machine intelligence toward greater autonomy and creative problem-solving. A notable example is *Move 37* by the *AlphaGo* model, which has been interpreted as one of the first demonstrations of out-of-distribution reasoning by a deep learning system [124]. Future research is to continue to explore training paradigms that enable models to exhibit both (i) high-quality interpolation and (ii) desirable extrapolation capabilities, while keeping the development of corresponding evaluation mechanisms in mind to ensure comparison, alignment, and interpretability of the modeling capabilities.

Current evaluation methods for generative models, commonly based on fidelity, diversity, or similarity to training data, are prone to become insufficient for capturing the higher-order capabilities such as reasoning, concept composition, and creative synthesis, as discussed in this chapter. To date, conventional metrics are by design not developed to assess whether a model can generalize beyond its training distribution or produce outputs with interpolated, semantic, logical, or epistemic depth. As generative models increasingly aim to support scientific reasoning and innovation, the need emerges to devise novel frameworks to analyze and optimize such models beyond the surface-level realism of their generated data. Future work is to explore extensions of existing evaluation paradigms not only for ensuring robustness when models encounter previously unseen data, such as novel imaging distributions, but also for enabling creative problem-solving and alignment of generative models with their intended objectives. Once corresponding evaluation mechanisms are well-informed,

7.4. The Intricacies of Evaluating Generative Models

interpretable, and, ideally, also optimizable, they can serve as a foundation for designing trustworthy and purpose-aligned generative artificial intelligence.

Chapter 8

Conclusion

This thesis explored and advanced generative deep learning methods to address critical limitations in cancer imaging, focusing on issues related to data availability, patient privacy, and diagnostic capabilities of deep learning models. Cancer imaging, particularly in fields like breast MRI, is often constrained by the scarcity of high-quality, annotated datasets due to privacy concerns and the labor-intensive nature of manual labeling. Traditional reliance on such data poses challenges for developing robust AI models that can generalize across diverse patient populations and imaging protocols. To address these challenges, this work investigated several approaches, including the use of conditional GANs and multi-conditional latent diffusion models, for synthesizing realistic cancer imaging data. These methods offer alternatives to conventional data acquisition by generating synthetic data that retains diagnostic utility, enabling AI models to learn from diverse and representative datasets without compromising patient privacy.

First, this thesis presented a large-scale survey and analysis of generative models applied in cancer imaging, identifying unmet needs and guiding further developments. From this analysis, the SynTRUST framework was introduced to systematically evaluate the trustworthiness and maturity of cancer image synthesis studies, setting a benchmark for the quality and reliability of synthetic data used in clinical research. Secondly a research study was presented that defined the desirable characteristics of synthetic data in a concise framework, while also developing conditional GANs for simulating dynamic contrast-enhanced (DCE) breast MRI sequences, generating high-quality images without the use of physical contrast agents. This approach not only circumvents health risks associated with contrast agents but also provides robust synthetic training data that

can be applied to unsupervised tumor detection, localization, and segmentation, thereby addressing the cancer imaging challenge of data and annotation scarcity. To this end, further methodological research proposed a multi-conditional latent diffusion model, which effectively translates non-contrast-enhanced MRI into dynamic, time-dependent synthetic DCE-MRI, allowing for improved tumor localization and presenting a first study analyzing the prediction of tumor contrast enhancement kinetics using diffusion models. This study further introduces the Fréchet Radiomics Distance (FRD), an interpretable and domain-specific quality metric for synthetic data, which focuses on clinical utility of synthetic imaging data by capturing the biomarker variability between synthetic and real datasets for both 2D and 3D imaging data.

A further contribution of this thesis is the development of a malignancy-conditioned GAN (MCGAN), designed to generate privacy-preserving mammography data that allows training of diagnostic models without access to real patient information. Experimental results demonstrated that MCGAN data, particularly when combined with differential privacy techniques, notably improves the performance of cancer classification models in privacy-sensitive scenarios. To facilitate broader access and collaboration, the thesis introduced the medigan library, a platform for sharing pretrained generative models across the research community, enabling accessible synthetic data generation for various imaging modalities including radiology and endoscopy domains. medigan includes comprehensive evaluations of its models including an evaluation of the variability of evaluation metrics like the Fréchet Inception Distance (FID) based on post-processing operations such as image resizing. These assessments performed through both general and radiology-specific feature extractors, ensuring that the synthetic data is validated in a multi-faceted manner to approximate clinical utility. Additionally, the FID ratio is proposed and included in this analysis, where the FID ratio considers the naturally occurring variation in the respective real dataset to better assess its corresponding synthetic dataset.

The advancements presented in this thesis lay a scientific foundation for generative models as pivotal tools in cancer imaging, supporting more accurate diagnostic interventions, enhancing patient privacy, and promoting equitable healthcare access. By addressing limitations in data availability and variability, this work brings the field closer to integrating synthetic data and generative models in real-world clinical settings. As generative deep learning models evolve, they offer a pathway toward data-driven healthcare solutions that improve diagnostic robustness, enable secure data sharing across institutions, and reduce reliance on traditional, patient-derived datasets. This thesis thus positions generative models not merely as supplementary tools but as

core components in the future of medical artificial intelligence, enhancing diagnostic outcomes while ensuring patient privacy. To this end, it contributes to a future where deep learning can be applied responsibly and effectively in oncology, advancing both individual patient care and the broader landscape of data-driven medical research.

Bibliography

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- [2] Abeer H Abdelhafez, Benjamin C Musall, Beatriz E Adrada, KennethR Hess, Jong Bum Son, Ken-Pin Hwang, Rosalind P Candelaria, Lumarie Santiago, Gary J Whitman, Huong T Le-Petross, et al. Tumor necrosis by pretreatment breast mri: association with neoadjuvant systemic therapy (nast) response in triple-negative breast cancer (tnbc). *Breast cancer research and treatment*, 185:1–12, 2021.
- [3] Leila Abdelrahman, Manal Al Ghamdi, Fernando Collado-Mesa, and Mohamed Abdel-Mottaleb. Convolutional neural networks for breast cancer detection in mammography: A survey. *Computers in Biology and Medicine*, 131(January):104248, 2021.
- [4] Sharmin Afrose, Wenjia Song, Charles B Nemeroff, Chang Lu, and Danfeng Yao. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications medicine*, 2(1):111, 2022.
- [5] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.
- [6] Marta Buetas Arcas, Richard Osuala, Karim Lekadir, and Oliver Díaz. Mitigating annotation shift in cancer classification using single-image generative models. In *17th International Workshop on Breast Imaging (IWBI 2024)*, volume 13174, pages 512–523. SPIE, 2024.
- [7] Chad A Arledge, Alan H Zhao, Umit Topaloglu, and Dawen Zhao. Dynamic contrast enhanced mri mapping of vascular permeability for evaluation of breast cancer neoadjuvant chemotherapy response using image-to-image conditional generative adversarial networks. *medRxiv*, 2024.
- [8] Blair Attard-Frost and David Gray Widder. The ethics of ai value chains. *arXiv preprint arXiv:2307.16787*, 2023.

Bibliography

- [9] Brendan Avent, Aleksandra Korolova, David Zeber, Torgeir Hovden, and Benjamin Livshits. {BLENDER}: Enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 747–764, 2017.
- [10] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023.
- [11] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerooni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.
- [12] S Bahri, J-H Chen, HJ Yu, A Kuzucan, O Nalcioğlu, and M-Y Su. Can dynamic contrast-enhanced mri (dce-mri) predict tumor recurrence and lymph node status in patients with breast cancer? *Annals of oncology*, 19(4):822–824, 2008.
- [13] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.
- [14] Andrea Barucci and Emanuele Neri. Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(13):2941–2943, 2020.
- [15] Björn Bebensee. Local differential privacy: a tutorial. *arXiv preprint arXiv:1907.11908*, 2019.
- [16] Y. Bengio and Yann Lecun. Convolutional networks for images, speech, and time-series. 11 1997.
- [17] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Diffusion models with implicit guidance for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–220. Springer, 2024.
- [18] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Evaluating normative representation learning in generative ai for robust anomaly detection in brain imaging. *Nature Communications*, 16(1):1624, 2025.
- [19] Eyal Bercovich and Marcia C Javitt. Medical imaging: from roentgen to the digital revolution, and beyond. *Rambam Maimonides medical journal*, 9(4), 2018.
- [20] Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtaash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

-
- [21] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.
 - [22] Helen Blumen, Kathryn Fitch, and Vincent Polkus. Comparison of treatment costs for breast cancer, by tumor stage and type of service. *American health & drug benefits*, 9(1):23, 2016.
 - [23] Margaret A Boden. Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2):347–356, 1998.
 - [24] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
 - [25] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.
 - [26] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.
 - [27] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.
 - [28] Adrian P Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, 8(1):171–182, 2017.
 - [29] Robert Brünjes and Thilo Hofmann. Anthropogenic gadolinium in freshwater and drinking water systems. *Water research*, 182:115966, 2020.
 - [30] Marco Caballo, Wendelien BG Sanderink, Luyi Han, Yuan Gao, Alexandra Athanasiou, and Ritse M Mann. Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI. *Journal of Magnetic Resonance Imaging*, 57(1):97–110, 2023.
 - [31] Víctor M Campello, Tian Xia, Xiao Liu, Pedro Sanchez, Carlos Martín-Isla, Steffen E Petersen, Santi Seguí, Sotirios A Tsaftaris, and Karim Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9:983091, 2022.
 - [32] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.
 - [33] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

Bibliography

- [34] Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, and Dimitris N Metaxas. Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13856–13866, 2020.
- [35] Yu-Cherng Channing Chang, Ellen Ackerstaff, Yohann Tschudi, Bryan Jimenez, Warren Foltz, Carl Fisher, Lothar Lilge, HyungJoon Cho, Sean Carlin, Robert J Gillies, et al. Delineation of tumor habitats based on dynamic contrast enhanced mri. *Scientific reports*, 7(1):9746, 2017.
- [36] Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abidin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.
- [37] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.
- [38] Xiaoran Chen, Nick Pawlowski, Ben Glocker, and Ender Konukoglu. Normative ascent with local gaussians for unsupervised lesion detection. *Medical Image Analysis*, 74:102208, 2021.
- [39] Liuquan Cheng and Xiru Li. Breast magnetic resonance imaging: kinetic curve assessment. *Gland surgery*, 2(1):50, 2013.
- [40] Eun Jung Choi, HyeMi Choi, Sin Ae Choi, and Ji Hyun Youk. Dynamic contrast-enhanced breast magnetic resonance imaging for the prediction of early and late recurrences in breast cancer. *Medicine*, 95(48):e5330, 2016.
- [41] Dan CireAan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32:333–338, 2012.
- [42] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.
- [43] Gary S Collins, Karel GM Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, et al. Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.
- [44] Luca Colombo, Alessandro Falcetta, and Manuel Roveri. Training encrypted neural networks on encrypted data with fully homomorphic encryption. In *Proceedings of the 12th Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, pages 64–75, 2023.

-
- [45] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024.
 - [46] Deniz Daum, Richard Osuala, Anneliese Riess, Georgios Kaissis, Julia A Schnabel, and Maxime Di Folco. On differentially private 3d medical image synthesis with controllable latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 139–149. Springer, 2024.
 - [47] Supratim De, Loubna Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.
 - [48] Helena M Dekker, Gerard J Stroomberg, Aart J Van der Molen, and Mathias Prokop. Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents. *Insights into Imaging*, 15(1):1–14, 2024.
 - [49] Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, and Fredrik Strand. Artificial intelligence for breast cancer detection in screening mammography in sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, 5(10):e703–e711, 2023.
 - [50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
 - [51] Jacob Devlin, Ming-Wei Chang, and Kenton Lee. Google, kt, language, ai: Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
 - [52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
 - [53] Oliver Diaz, Kaisar Kushibar, Richard Osuala, Akis Linardos, Lidia Garrucho, Laura Igual, Petia Radeva, Fred Prior, Polyxeni Gkontra, and Karim Lekadir. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica Medica*, 83:25–37, 2021.
 - [54] Apple Differential Privacy Team. Learning with Privacy at Scale. 2017.
 - [55] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Bibliography

- [56] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- [57] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [58] Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto dos Santos. Automation bias in mammography: the impact of artificial intelligence bi-rads suggestions on reader performance. *Radiology*, 307(4):e222176, 2023.
- [59] Joshua Durso-Finley, Berardino Barile, Jean-Pierre Falet, Douglas L Arnold, Nick Pawlowski, and Tal Arbel. Probabilistic temporal prediction of continuous disease trajectories and treatment effects using neural sdes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 400–410. Springer, 2024.
- [60] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [61] Riham H El Khouli, Katarzyna J Macura, Michael A Jacobs, Tarek H Khalil, Ihab R Kamel, Andrew Dwyer, and David A Bluemke. Dynamic contrast-enhanced mri of the breast: quantitative method for kinetic curve type assessment. *American Journal of Roentgenology*, 193(4):W295–W300, 2009.
- [62] Joann G Elmore, Carolyn K Wells, Carol H Lee, Debra H Howard, and Alvan R Feinstein. Variability in radiologists’ interpretations of mammograms. *New England Journal of Medicine*, 331(22):1493–1499, 1994.
- [63] Masahiro Endo. History of medical physics. *Radiological Physics and Technology*, 14(4):345–357, 2021.
- [64] Úlfar Erlingsson, Vasył Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.
- [65] European Medicines Agency (EMA). EMA’s final opinion confirms restrictions on use of linear gadolinium agents in body scans. <https://www.ema.europa.eu/>, 2023. Online; accessed 06 August 2023.
- [66] European Parliament and Council of European Union. General Data Protection Regulation (GDPR), REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Online at <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679/>, 2018.

- [67] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5):E359–E386, 2015.
- [68] Virginia Fernandez, Pedro Sanchez, Walter Hugo Lopez Pinaya, Grzegorz Jancenków, Sotirios A. Tsaftaris, and M. Jorge Cardoso. Privacy distillation: Reducing re-identification risk of diffusion models. In *Deep Generative Models: Third MICCAI Workshop, DGM4MICCAI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings*, page 3–13, Berlin, Heidelberg, 2024. Springer-Verlag.
- [69] Barbara M Fischer, Minna WB Olsen, Carsten D Ley, Thomas L Klausen, Jann Mortensen, Liselotte Højgaard, and Paul EG Kristjansen. How few cancer cells can be detected by positron emission tomography? A frequent question addressed by an in vitro study. *European journal of nuclear medicine and molecular imaging*, 33(6):697–702, 2006.
- [70] Stefan M Fischer, Lina Felsner, Richard Osuala, Johannes Kiechle, Daniel M Lang, Jan C Peeken, and Julia A Schnabel. Progressive growing of patch size: Resource-efficient curriculum learning for dense prediction tasks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–520. Springer, 2024.
- [71] John V Frangioni. New technologies for human cancer imaging. *Journal of clinical oncology*, 26(24):4012, 2008.
- [72] Kunihiro Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.
- [73] Kunihiro Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.
- [74] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- [75] Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, pages 1–10, 2025.
- [76] Zeyu Gan and Yong Liu. Towards a theoretical understanding of synthetic data in llm post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*, 2024.

Bibliography

- [77] Lidia Garrucho, Eve Delege, Richard Osuala, Dimitri Kessler, Kaisar Kushibar, Oliver Díaz, Karim Lekadir, and Laura Igual. Fat-suppressed breast mri synthesis for domain adaptation in tumour segmentation. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pages 202–211. Springer, 2024.
- [78] Lidia Garrucho, Kaisar Kushibar, Socayna Jouide, Oliver Diaz, Laura Igual, and Karim Lekadir. Domain generalization in deep learning-based mass detection in mammography: A large-scale multi-center study. *arXiv preprint arXiv:2201.11620*, 2022.
- [79] Lidia Garrucho, Kaisar Kushibar, Richard Osuala, Oliver Diaz, Alessandro Catanese, Javier Del Riego, Maciej Bobowicz, Fredrik Strand, Laura Igual, and Karim Lekadir. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12:7424, 2023.
- [80] Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździwicz, Maria-Laura Cosaka, Pasant M. Abo-Elhoda, Sara W. Tantawy, Shorouq S. Sakrana, Norhan O. Shawky-Abdelfatah, Amr Muhammad Abdo-Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E. Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpınar, Oğuz Lafcı, Ritse Mann, Carlos Martín-Isla, Fred Prior, Kostas Marias, Martijn P.A. Starmans, Fredrik Strand, Oliver Díaz, Laura Igual, and Karim Lekadir. A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv preprint*, 2024.
- [81] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.
- [82] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.
- [83] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [84] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [85] Mark S Graham, Walter Hugo Lopez Pinaya, Paul Wright, Petru-Daniel Tudosi, Yee H Mah, James T Teo, H Rolf Jäger, David Werring, Parashkev Nachev, Sebastien Ourselin, et al. Unsupervised 3d out-of-distribution detection with

- latent diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 446–456. Springer, 2023.
- [86] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.07758*, 2022.
 - [87] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.
 - [88] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.
 - [89] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.
 - [90] Nathan R Hill, Samuel T Fatoba, Jason L Oke, Jennifer A Hirst, Christopher A O’Callaghan, Daniel S Lasserson, and FD Richard Hobbs. Global prevalence of chronic kidney disease—a systematic review and meta-analysis. *PloS one*, 11(7):e0158765, 2016.
 - [91] Eishu Hirata and Erik Sahai. Tumor microenvironment and differential responses to therapy. *Cold Spring Harbor perspectives in medicine*, 7(7):a026781, 2017.
 - [92] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
 - [93] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.
 - [94] Erich P Huang, James PB O’Connor, Lisa M McShane, Maryellen L Giger, Philippe Lambin, Paul E Kinahan, Eliot L Siegel, and Lalitha K Shankar. Criteria for the translation of radiomics into clinically useful tests. *Nature reviews Clinical oncology*, 20(2):69–82, 2023.
 - [95] Christopher H Hunt, Robert P Hartman, and Gina K Hesley. Frequency and severity of adverse effects of iodinated and gadolinium contrast materials: retrospective review of 456,930 doses. *American Journal of Roentgenology*, 193(4):1124–1127, 2009.
 - [96] Erik Johannes Husom, Arda Goknil, Lwin Khin Shar, and Sagar Sen. The price of prompting: Profiling energy use in large language models inference. *arXiv preprint arXiv:2407.16893*, 2024.
 - [97] Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo JWL Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 17(12):771–781, 2020.

Bibliography

- [98] Jean-Marc Idée, Marc Port, Isabelle Raynal, Michel Schaefer, Soizic Le Greneur, and Claire Corot. Clinical and biological consequences of transmetallation induced by contrast agents for magnetic resonance imaging: a review. *Fundamental & clinical pharmacology*, 20(6):563–576, 2006.
- [99] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [100] Jason N Itri, Rafel R Tappouni, Rachel O McEachern, Arthur J Pesch, and Sohil H Patel. Fundamentals of diagnostic error in imaging. *Radiographics*, 38(6):1845–1865, 2018.
- [101] Ahmedin Jemal, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011.
- [102] Changzhe Jiao, Diane Ling, Shelly Bian, April Vasantachart, Karen Cheng, Shahil Mehta, Derrick Lock, Zhenyu Zhu, Mary Feng, Horatio Thomas, et al. Contrast-enhanced liver magnetic resonance image synthesis using gradient regularized multi-modal multi-discrimination sparse attention fusion gan. *Cancers*, 15(14):3544, 2023.
- [103] Ming-Zhu Jin and Wei-Lin Jin. The updated landscape of tumor microenvironment and drug repurposing. *Signal transduction and targeted therapy*, 5(1):166, 2020.
- [104] Jennifer Viberg Johansson, Karin Dembrower, Fredrik Strand, and Åsa Grauman. Women’s perceptions and attitudes towards the use of ai in mammography in sweden: a qualitative interview study. *BMJ open*, 14(2):e084014, 2024.
- [105] Smriti Joshi, Richard Osuala, Lidia Garrucho, Apostolia Tsirikoglou, Javier del Riego, Katarzyna Gwoździwicz, Kaisar Kushibar, Oliver Diaz, and Karim Lekadir. Leveraging epistemic uncertainty to improve tumour segmentation in breast mri: an exploratory analysis. In *Medical Imaging 2024: Image Processing*, volume 12926, pages 292–300. SPIE, 2024.
- [106] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
- [107] Thorsten Kalb, Kaisar Kushibar, Celia Cintas, Karim Lekadir, Oliver Diaz, and Richard Osuala. Revisiting skin tone fairness in dermatological lesion classification. In *Workshop on Clinical Image-Based Procedures*, pages 246–255. Springer, 2023.
- [108] Tomonori Kanda, Kazunari Ishii, Hiroki Kawaguchi, Kazuhiro Kitajima, and Daisuke Takenaka. High signal intensity in the dentate nucleus and globus pallidus

- on unenhanced t1-weighted mr images: relationship with increasing cumulative dose of a gadolinium-based contrast material. *Radiology*, 270(3):834–841, 2014.
- [109] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
- [110] Narendra N Khanna, Mahesh A Maindarkar, Vijay Viswanathan, Jose Fernandes E Fernandes, Sudip Paul, Mrinalini Bhagawati, Puneet Ahluwalia, Zoltan Ruzsa, Aditya Sharma, Raghu Kolluri, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. In *Healthcare*, volume 10, page 2493. MDPI, 2022.
- [111] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [112] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [113] Nicholas Konz, Richard Osuala, Preeti Verma, Yuwen Chen, Hanxue Gu, Haoyu Dong, Yaqian Chen, Andrew Marshall, Lidia Garrucho, Kaisar Kushibar, Daniel M. Lang, Gene S. Kim, Lars J. Grimm, John M. Lewin, James S. Duncan, Julia A. Schnabel, Oliver Diaz, Karim Lekadir, and Maciej A. Mazurowski. Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. 2025.
- [114] Csaba P Kovesdy. Epidemiology of chronic kidney disease: an update 2022. *Kidney international supplements*, 12(1):7–11, 2022.
- [115] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [116] Karoline B Kuchenbaecker, John L Hopper, Daniel R Barnes, Kelly-Anne Phillips, Thea M Mooij, Marie-José Roos-Blom, Sarah Jervis, Flora E Van Leeuwen, Roger L Milne, Nadine Andrieu, et al. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.
- [117] Christiane Katharina Kuhl, Peter Mielcareck, Sven Klaschik, Claudia Leutner, Eva Wardelmann, Jurgen Gieseke, and Hans H Schild. Dynamic breast mr imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–110, 1999.
- [118] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David

Bibliography

- Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.
- [119] Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–524. Springer, 2022.
- [120] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- [121] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.
- [122] Daniel M Lang, Eli Schwartz, Cosmin I Bercea, Raja Giryes, and Julia A Schnabel. Multispectral 3d masked autoencoders for anomaly detection in non-contrast enhanced breast mri. In *MICCAI Workshop on Cancer Prevention through Early Detection*, pages 55–67. Springer, 2023.
- [123] Paul C Lauterbur. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *nature*, 242(5394):190–191, 1973.
- [124] Neil Leach. In the mirror of ai: what is creativity? *Architectural Intelligence*, 1(1):15, 2022.
- [125] Y LeCun. Generalization and network design strategies. *Connections in Perspective*, 1989.
- [126] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [127] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [128] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.
- [129] Karim Lekadir, Aasa Feragen, Abdul Joseph Fofanah, Alejandro F Frangi, Alena Buyx, Anais Emelie, Andrea Lara, Antonio R Porras, An-Wen Chan, Arcadi Navarro, et al. Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *arXiv preprint arXiv:2309.12325*, 2023.

-
- [130] Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, Georgios Kaissis, Gianna Tsakou, Irène Buvat, Jayashree Kalpathy-Cramer, John Mongan, Julia A Schnabel, Kaisar Kushibar, Katrine Riklund, Kostas Marias, Lameck M Amugongo, Lauren A Fromont, Lena Maier-Hein, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, M Jorge Cardoso, Maciej Bobowicz, Mahsa Shabani, Manolis Tsiknakis, Maria A Zuluaga, Marie-Christine Fritzsche, Marina Camacho, Marius George Linguraru, Markus Wenzel, Marleen De Bruijne, Martin G Tolsgaard, Melanie Goisauf, Mónica Cano Abadía, Nikolaos Papanikolaou, Noussair Lazrak, Oriol Pujol, Richard Osuala, Sandy Napel, Sara Colantonio, Smriti Joshi, Stefan Klein, Susanna Aussó, Wendy A Rogers, Zohaib Salahuddin, and Martijn P A Starmans. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388, 2025.
- [131] Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, et al. Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv preprint arXiv:2109.09658*, 2021.
- [132] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025.
- [133] Jun Li, Che Liu, Wenjia Bai, Rossella Arcucci, Cosmin I Bercea, and Julia A Schnabel. Enhancing abnormality grounding for vision language models with knowledge descriptions. *arXiv preprint arXiv:2503.03278*, 2025.
- [134] Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:2304.03271*, 2023.
- [135] Xia Li, Lori R Arlinghaus, Gregory D Ayers, A Bapsi Chakravarthy, Richard G Abramson, Vandana G Abramson, Nkiruka Atuegwu, Jaime Farley, Ingrid A Mayer, Mark C Kelley, et al. Dce-mri analysis methods for predicting the response of breast cancer to neoadjuvant chemotherapy: Pilot study findings. *Magnetic resonance in medicine*, 71(4):1592–1602, 2014.
- [136] Zhuohang Li, Andrew Lowy, Jing Liu, Toshiaki Koike-Akino, Kieran Parsons, Bradley Malin, and Ye Wang. Analyzing inference privacy risks through gradients in machine learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3466–3480, 2024.
- [137] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bibliography

- [138] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [139] Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. *arXiv preprint arXiv:2412.01720*, 2024.
- [140] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [141] Jessica Lohrke, Thomas Frenzel, Jan Endrikat, Filipe Caseiro Alves, Thomas M Grist, Meng Law, Jeong Min Lee, Tim Leiner, Kun-Cheng Li, Konstantin Nikolaou, et al. 25 years of contrast-enhanced mri: developments, current challenges and future perspectives. *Advances in therapy*, 33:1–28, 2016.
- [142] Cesar Augusto Fontanillo López. On the legal nature of synthetic data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.
- [143] Miguel G Lopez, Naimy Posada, Daniel C Moura, Raúl Ramos Pollán, José M Franco Valiente, César Suárez Ortega, Manuel Solar, Guillermo Diaz-Herrero, IMAP Ramos, Joana Loureiro, et al. BCDR: a breast cancer digital repository. In *15th International conference on experimental mechanics*, volume 1215, 2012.
- [144] Peter Marckmann, Lone Skov, Kristian Rossen, Anders Dupont, Mette Brimnes Damholt, James Goya Heaf, and Henrik S Thomsen. Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging. *Journal of the American Society of Nephrology*, 17(9):2359–2362, 2006.
- [145] Carlos Martin-Isla, Victor M Campello, Cristian Izquierdo, Zahra Raisi-Estabragh, Bettina Baekler, Steffen E Petersen, and Karim Lekadir. Image-based cardiac diagnosis with machine learning: a review. *Frontiers in cardiovascular medicine*, 7:1, 2020.
- [146] G McCreadie and TB Oliver. Eight CT lessons that we learned the hard way: an analysis of current patterns of radiological error and discrepancy with particular emphasis on CT. *Clinical radiology*, 64(5):491–499, 2009.
- [147] Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology*, 22(9):1191–1198, 2015.

-
- [148] Robert J McDonald, Jeffrey C Weinreb, and Matthew S Davenport. Symptoms associated with gadolinium exposure (sage): a suggested term. *Radiology*, 302(2):270–273, 2022.
- [149] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.
- [150] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [151] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, page e210315, 2022.
- [152] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [153] Torgeir Mo, Siri Helene Bertelsen Brandal, Oliver Marcel Geier, Olav Engebråten, Line Brennhaug Nilsen, Vessela N Kristensen, Knut Håkon Hole, Tord Hompland, Thomas Fleischer, and Therese Seierstad. Mri assessment of changes in tumor vascularization during neoadjuvant anti-angiogenic treatment in locally advanced breast cancer patients. *Cancers*, 15(18):4662, 2023.
- [154] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.
- [155] James Muldoon, Callum Cant, Mark Graham, and Funda Ustek Spilda. The poverty of ethical ai: impact sourcing and ai supply chains. *AI & society*, pages 1–15, 2023.
- [156] James Muldoon, Callum Cant, Boxi Wu, and Mark Graham. A typology of artificial intelligence data work. *Big data & society*, 11(1):20539517241232632, 2024.
- [157] Gustav Müller-Franzes, Luisa Huck, Maike Bode, Sven Nebelung, Christiane Kuhl, Daniel Truhn, and Teresa Lemainque. Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast mri. *European Radiology Experimental*, 8(1):53, 2024.
- [158] Gustav Müller-Franzes, Luisa Huck, Soroosh Tayebi Arasteh, Firas Khader, Tianyu Han, Volkmar Schulz, Ebba Dethlefsen, Jakob Nikolas Kather, Sven Nebelung, Teresa Nolte, et al. Using Machine Learning to Reduce the Need

Bibliography

- for Contrast Agents in Breast MRI through Synthetic Images. *Radiology*, 307(3):e222211, 2023.
- [159] Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [160] Steven A Narod and William D Foulkes. Brca1 and brca2: 1994 and beyond. *Nature Reviews Cancer*, 4(9):665–676, 2004.
- [161] National Cancer Institute. Brca gene changes: Cancer risk and genetic testing. <https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet>, 2024. Accessed: 2025-02-08.
- [162] David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the “Big Three”. *Diagnosis*, 8(1):67–84, 2021.
- [163] Nghi C Nguyen, Theodore T Molnar, Lucas G Cummin, and Emanuel Kanal. Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine administrations: a retrospective analysis. *Radiology*, 296(1):122–130, 2020.
- [164] Larry Norton, Richard Simon, Harmar D Brereton, and Arthur E Bogden. Predicting the course of Gompertzian growth. *Nature*, 264(5586):542–545, 1976.
- [165] Cyprian Olchow, Kamil Cebulski, Mateusz Łasecki, Radosław Chaber, Anna Olchow, Krzysztof Kałwak, and Urszula Zaleska-Dorobisz. The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-a systematic review. *PloS one*, 12(2):e0171704, 2017.
- [166] Richard Osuala, Smriti Joshi, Apostolia Tsirikoglou, Lidia Garrucho, Walter HL Pinaya, Oliver Diaz, and Karim Lekadir. Pre-to post-contrast breast mri synthesis for enhanced tumour segmentation. In *Medical Imaging 2024: Image Processing*, volume 12926, pages 226–237. SPIE, 2024.
- [167] Richard Osuala, Jieyi Li, and Ognjen Arandjelovic. Bringing modern machine learning into clinical practice through the use of intuitive visualization and human-computer interaction. *Augmented Human Research*, 4(1):3, 2019.
- [168] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.
- [169] Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. In *Medical Imaging with Deep Learning*, 2018.

-
- [170] Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 705–714. Springer, 2022.
- [171] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022.
- [172] Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5):1113–1122, 2023.
- [173] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [174] Zahra Raisi-Estabragh, Polyxeni Gkontra, Akshay Jaggi, Jackie Cooper, João Augusto, Anish N Bhuva, Rhodri H Davies, Charlotte H Manisty, James C Moon, Patricia B Munroe, et al. Repeatability of cardiac magnetic resonance radiomics: a multi-centre multi-vendor test-retest study. *Frontiers in Cardiovascular Medicine*, 7:586236, 2020.
- [175] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.
- [176] Beatriu Reig, Alana A Lewin, Linda Du, Laura Heacock, Hildegard K Toth, Samantha L Heller, Yiming Gao, and Linda Moy. Breast mri for evaluation of response to neoadjuvant therapy. *Radiographics*, 41(3):665–679, 2021.
- [177] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.
- [178] Mohammad H Rezazade Mehrizi, Ferdinand Mol, Marcel Peter, Erik Ranschaert, Daniel Pinto Dos Santos, Ramin Shahidi, Mansoor Fatehi, and Thomas Dratsch. The impact of ai suggestions on radiologists’ decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Scientific Reports*, 13(1):9230, 2023.
- [179] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [180] Justyna Rogowska, Ewa Olkowska, Wojciech Ratajczyk, and Lidia Wolska. Gadolinium as a new emerging contaminant of aquatic environments. *Environmental toxicology and chemistry*, 37(6):1523–1534, 2018.

Bibliography

- [181] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2021.
- [182] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [183] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [184] Solve Sæbø and Helge Brovold. On the stochastics of human and artificial creativity. *arXiv preprint arXiv:2403.06996*, 2024.
- [185] A Saha, M R Harowicz, L J Grimm, J Weng, E H Cain, C E Kim, S V Ghate, R Walsh, and M A Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. *The Cancer Imaging Archive*, 2021.
- [186] Mattie Salim, Erik Wählin, Karin Dembrower, Edward Azavedo, Theodoros Foukakis, Yue Liu, Kevin Smith, Martin Eklund, and Fredrik Strand. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA oncology*, 6(10):1581–1588, 2020.
- [187] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE, 2023.
- [188] Thomas Schaffter, Diana SM Buist, Christoph I Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA network open*, 3(3):e200265–e200265, 2020.
- [189] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.
- [190] Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *International conference on discovery science*, pages 26–38. Springer, 2007.

-
- [191] Hannes Schreiter, Jessica Eberle, Lorenz A Kapsner, Dominique Hadler, Sabine Ohlmeyer, Ramona Erber, Julius Emons, Frederik B Laun, Michael Uder, Evelyn Wenkel, et al. Virtual dynamic contrast enhanced breast mri using 2d u-net architectures. *medRxiv*, pages 2024–08, 2024.
- [192] Raghavendra Selvan, Nikhil Bhagwat, Lasse F Wolff Anthony, Benjamin Kanding, and Erik B Dam. Carbon footprint of selecting and training deep learning models for medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 506–516. Springer, 2022.
- [193] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
- [194] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [195] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- [196] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Enhancing ai systems with agentic workflows patterns in large language model. In *2024 IEEE World AI IoT Congress (AIIoT)*, pages 527–532. IEEE, 2024.
- [197] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [198] Grzegorz Skorupko, Richard Osuala, Zuzanna Szafranowska, Kaisar Kushibar, Nay Aung, Steffen E Petersen, Karim Lekadir, and Polyxeni Gkontra. Debi-asing cardiac imaging with controlled latent diffusion models. *arXiv preprint arXiv:2403.19508*, 2024.
- [199] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [200] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [201] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Bibliography

- [202] Martijn Starmans, Sebastian R van der Voort, Thomas Phil, Milea JM Timbergen, Melissa Vos, Guillaume A Padmos, Wouter Kessels, David Hanff, Dirk J Grunhagen, Cornelis Verhoef, et al. Reproducible radiomics through automated machine learning validated on twelve clinical applications. *arXiv preprint arXiv:2108.08618*, 2021.
- [203] Ruolin Su, Xiao Liu, and Sotirios A Tsaftaris. Why patient data cannot be easily forgotten? In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 632–641. Springer, 2022.
- [204] Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D Cox, and Akash Srivastava. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*, 2024.
- [205] Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022.
- [206] Zuzanna Szafranowska, Richard Osuala, Bennet Breier, Kaisar Kushibar, Karim Lekadir, and Oliver Diaz. Sharing generative models instead of private data: a simulation study on mammography patch classification. In *16th International Workshop on Breast Imaging (IWBI2022)*, volume 12286, pages 169–177. SPIE, 2022.
- [207] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [208] Xiuli Tao, Lvhuang Wang, Zhouguang Hui, Li Liu, Feng Ye, Ying Song, Yu Tang, Yu Men, Tryphon Lambrou, Zihua Su, et al. Dce-mri perfusion and permeability parameters as predictors of tumor response to ccrt in patients with locally advanced nslc. *Scientific reports*, 6(1):35569, 2016.
- [209] Rajat Thawani, Lina Gao, Ajay Mohinani, Alina Tudorica, Xin Li, Zahi Mitri, and Wei Huang. Quantitative dce-mri prediction of breast cancer recurrence following neoadjuvant chemotherapy: a preliminary study. *BMC Medical Imaging*, 22(1):182, 2022.
- [210] Yee Liang Thian, Angela M Riddell, and Dow-Mu Koh. Liver-specific agents for contrast-enhanced mri: role in oncological imaging. *Cancer imaging*, 13(4):567, 2013.
- [211] Adrian MK Thomas and Arpan K Banerjee. *The history of radiology*. OUP Oxford, 2013.
- [212] U.S. Congress. Health insurance portability and accountability act of 1996. <https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf>, 1996.

-
- [213] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
- [214] Janna E van Timmeren, Ralph TH Leijenaar, Wouter van Elmpt, Jiazhou Wang, Zhen Zhang, André Dekker, and Philippe Lambin. Test–retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography*, 2(4):361, 2016.
- [215] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [216] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.
- [217] Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J Su, Zhun Deng, Michael Qizhe Xie, Hannah Brown, et al. Can ai be as creative as humans? *arXiv preprint arXiv:2401.01623*, 2024.
- [218] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [219] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [220] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [221] Christopher L Welle, Flavius F Guglielmo, and Sudhakar K Venkatesh. Mri of the liver: choosing the right contrast agent. *Abdominal radiology*, 45:384–392, 2020.
- [222] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.
- [223] MinJae Woo, Steven C Lowe, A Michael Devane, and Ronald W Gimbel. Intervention to reduce interobserver variability in computed tomographic measurement of cancer lesions among experienced radiologists. *Current problems in diagnostic radiology*, 2020.

Bibliography

- [224] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2024.
- [225] World Health Organization. Cancer, 2018.
- [226] World Health Organization. Cancer, 2024.
- [227] Shandong Wu, Wendie A Berg, Margarita L Zuley, Brenda F Kurland, Rachel C Jankowitz, Robert Nishikawa, David Gur, and Jules H Sumkin. Breast mri contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer. *Breast Cancer Research*, 18:1–10, 2016.
- [228] Ting Wu and Yun Dai. Tumor microenvironment and therapeutic response. *Cancer letters*, 387:61–68, 2017.
- [229] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anodpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.
- [230] Xiaodan Xing, Federico Felder, Yang Nan, Giorgos Papanastasiou, Simon Walsh, and Guang Yang. You don’t have to be perfect to be amazing: Unveil the utility of synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–22. Springer, 2023.
- [231] Runhua Xu, James BD Joshi, and Chao Li. Cryptonn: Training neural networks over encrypted data. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1199–1209. IEEE, 2019.
- [232] Guo-Ping Yan, Leslie Robinson, and Peter Hogg. Magnetic resonance imaging contrast agents: overview and perspectives. *Radiography*, 13:e5–e19, 2007.
- [233] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- [234] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [235] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

- [236] Tianyu Zhang, Luyi Han, Anna D'Angelo, Xin Wang, Yuan Gao, Chunyao Lu, Jonas Teuwen, Regina Beets-Tan, Tao Tan, and Ritse Mann. Synthesis of Contrast-Enhanced Breast MRI Using Multi-b-Value DWI-based Hierarchical Fusion Network with Attention Mechanism. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 79–88, Cham, 2023. Springer Nature Switzerland.