

UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S  
THESIS

---

**Synthetic training data generation from a  
single image for enhanced breast cancer  
diagnosis.**

---

*Author:*  
Marta BUETAS ARCAS

*Supervisors:*  
Oliver DÍAZ  
Richard OSUALA

*A thesis submitted in partial fulfillment of the requirements  
for the degree of MSc in Fundamental Principles of Data Science*

*in the*

Facultat de Matemàtiques i Informàtica

29th June 2023



UNIVERSITAT DE BARCELONA

*Abstract*

Facultat de Matemàtiques i Informàtica

MSc

**Synthetic training data generation from a single image for enhanced breast cancer diagnosis.**

by Marta BUETAS ARCAS

According to the World Health Organisation (WHO), breast cancer is one of the cancer types with a high prevalence worldwide. Deep-learning based computer-aided detection systems have shown promising potential in improving the curability and reducing mortality rates through early detection in mammography screening. Artificial Intelligence (AI) has become a popular tool in medicine, aiming to reduce costs and assist radiologists in decision-making processes. However, AI in cancer imaging presents significant challenges, including data access and privacy issues, as well as a scarcity of expert-annotated medical imaging. Motivated by these factors, this project aims to enhance the robustness and generalisability of breast cancer classification tools. The study focuses on obtaining a pre-biopsy result of suspicious areas in mammograms, providing a comprehensive assessment of lesion nature. It was observed that the classifier's performance for the malignant class was inferior to that of the other classes, and the tightness of the annotation mask around the lesion significantly influenced the classifier's performance. To improve the performance for malignant lesions, the study investigates data augmentation based in single image Generative Adversarial Network (SinGAN) to balance this underrepresented class. To the best of our knowledge, this project represents a novel investigation into the application of single-image generative models for breast cancer, addressing the challenge of expert annotation scarcity. Promising results were observed through the use of SinGAN-based data augmentation. The classification model, trained with SinGAN-augmented training data, demonstrated a higher area under the receiver operating characteristic (AUROC) for the malignant class ( $0.718 \pm 0.044$ ), compared to the same model without augmented data ( $0.677 \pm 0.076$ ). Furthermore, it was also identified an unexpected trend during the experiments. It was observed that using more SinGANs for data augmentation did not always result in a higher enhancement of performance. This project opens up new research possibilities through collaboration with healthcare experts. Its ultimate goal is to analyse and validate a mitigation strategy for improving robustness and, as such, trustworthiness of AI-based applications for adoption in the clinical workflow.



## *Acknowledgements*

Firstly, I would like to express my gratitude to my supervisors, Oliver and Richard. They have provided me with the opportunity and trusted me to develop this project, which I have greatly enjoyed and learned from. I would also like to thank my professors from Universitat de Barcelona, who have provided valuable and useful tools for the development of this work. Additionally, they have made this year of studies an enriching experience both academically and personally. I would also like to thank my classmates from the master's program and my family, who have been the best colleagues to share this journey with. Lastly, I extend my thanks to the reader of this report, for their time, interest, and any potential suggestions for improvement or even new contributions.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Challenges of Artificial Intelligence in medicine . . . . .	1
1.3 Contributions . . . . .	3
<b>2 Literature review</b>	<b>5</b>
<b>3 Methods and materials</b>	<b>9</b>
3.1 BCDR dataset . . . . .	9
3.2 Mammogram patch extraction . . . . .	9
3.2.1 Rationale . . . . .	9
3.2.2 Technique . . . . .	10
3.3 Analysis of the data . . . . .	13
3.4 Data augmentation with SinGAN . . . . .	16
3.4.1 SinGAN model architecture . . . . .	16
3.4.2 SinGAN implementation . . . . .	18
3.4.3 SiFID as evaluation metric . . . . .	20
3.5 Classification pipeline . . . . .	21
3.5.1 Evaluation of the classification . . . . .	22
3.5.2 Malignant class balancing . . . . .	23
3.5.3 Varying the zoom level group with which the classifier is trained and tested . . . . .	23
<b>4 Results and discussion</b>	<b>25</b>
4.1 Binary experiment . . . . .	25
4.2 Training at different zoom levels of the lesions . . . . .	26
4.3 Analysing classifier robustness against variations in the quality or ac- curacy of the annotation mask . . . . .	28
4.4 SinGAN-based data augmentation . . . . .	28
4.4.1 Generation and evaluation images with SinGAN models . . . . .	29
4.4.2 Improving baseline performance with SinGAN-based data aug- mentation . . . . .	30
4.4.3 Synthetic dataset generated from a different number of SinGAN models . . . . .	32
<b>5 Conclusions and further work</b>	<b>35</b>
<b>Bibliography</b>	<b>39</b>

<b>6</b>	<b>Appendix</b>	<b>41</b>
6.1	ResNet model graph . . . . .	41
6.2	ROC curves . . . . .	42
6.2.1	ROC curves of experiments computed in section 4.2 . . . . .	42
6.2.2	ROC curves of experiments computed in section 4.3 . . . . .	43
6.2.3	ROC curves of experiments computed in section 4.4.3 . . . . .	44



## Chapter 1

# Introduction

### 1.1 Motivation

Cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, according to *WHO: Cancer Fact Sheet, 2022*. The most common ones are breast, lung, colon and rectum and prostate cancers. These figures can be reduced when cases are detected and treated early. An early diagnosis increases the probabilities of survival, as it is more likely to respond to treatment. On the other hand, screening programmes help to identify findings suggestive of a specific cancer or pre-cancer before they have developed symptoms.

To detect and diagnose tumours, radiologists inspect, normally by visual assessment, medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound (US), X-ray mammography (MMG) or nuclear imaging (positron emission tomography, PET or single-photon emission computerised tomography, SPECT).

The present project is focused on breast cancer that, as stated before, is one of the cancer types with the highest prevalence in the world. According to *WHO: Breast Cancer Fact Sheet 2021*, in 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. WHO reports that breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life. The organisation also highlights that improvements in survival began in the 1980s in countries with early detection programmes combined with different modes of treatment to eradicate invasive disease.

### 1.2 Challenges of Artificial Intelligence in medicine

Deep-learning based computer-aided detection systems have shown promising potential in enhancing the curability and reducing mortality rates of breast cancer through early detection in mammography screening (MMG).

In general, Artificial Intelligence (AI) has become a popular tool in medicine to improve the performance of clinicians, to reduce costs and assist radiologists in the decision-making process. Despite technological and medical advances, the detection and diagnosis of breast cancer based on images continue to pose important challenges.

In this respect, AI models learn from data; thus, the amount and quality of medical data has a direct influence on the success of AI-based applications. The most

significant challenges of AI in cancer imaging can be resumed in: data scarcity, data access and privacy, data annotation and segmentation, cancer detection and diagnosis and tumour profiling and treatment (Osuala et al., 2023a).

Regarding data scarcity, it poses a significant challenge in developing AI tools for breast cancer detection. Scarcity of expert-annotated medical images often constrains Deep Learning based methods to be trained and evaluated on a small dataset coming from a single centre. Addressing this challenge involves establishing collaborations between healthcare institutions to pool together diverse and well-curated datasets. Additionally, techniques such as data augmentation could be employed to mitigate the impact of data scarcity and enhance the generalisability of AI models. In the development of this project, the impact of using synthetic data for both, data augmentation and addressing data imbalances was studied. Moreover, accessing scarce data becomes a more significant challenge when robust patient anonymisation is required.

One critical issue arising from data scarcity is class imbalance, which refers to the overrepresentation of certain types of data compared to others (Bi et al., 2019). One common manifestation of class imbalance is the imbalance of diagnostic labels because of a low prevalence of the disease in the population. For instance, in the context of breast cancer, the malignant label is often the minority class. Such imbalances can have a detrimental effect on a model's specificity or sensitivity, as the learned bias from the data distribution may impact its performance.

To tackle this problem, one approach is to address the imbalances by using synthetic images generated from models such as Generative Adversarial Networks (GANs), from Goodfellow et al., 2020. Hu et al., 2018 previously implemented GANs to generate underrepresented grades in a risk assessment scoring system for prostate cancer. In their paper, Szafranowska et al., 2022 proposed GANs as an alternative data-sharing method. They explored the concept of sharing trained generative models instead of the original private data, studying its potential in the context of mammography patch classification.

SinGAN (Shaham, Dekel and Michaeli, 2019), another promising framework, offers a solution to alleviate the challenge of data scarcity in cancer imaging. SinGAN generates multiple synthetic images based on a single training image, thereby enhancing the utility of each cancer imaging sample.

Integration of AI tools into radiology workflows, where they serve as decision support systems, can help improve accuracy, reduce false negatives, and provide second opinions, leading to more reliable and timely diagnoses. However, this involves developing robust Deep Learning models that can accurately detect and classify breast cancer lesions in order to achieve a wider adoption of this tools in clinical practice.

Addressing these challenges requires collaboration between clinicians, data scientists, policymakers and other stakeholders to establish robust data sharing mechanisms, develop ethical guidelines, and ensure the safe and effective implementation of AI tools in breast cancer detection and treatment. Trustworthiness and robustness are two of the keys to achieve a wider adoption of AI tools in medicine, as defined by the FUTURE-AI guidelines developed by Lekadir et al., 2021.

## 1.3 Contributions

The research is centered on analysing and enhancing the robustness and generalisability of a classification model for wider adoption in the clinical workflow. The classification task is to obtain a pre-biopsy result of breast lesions, which can assist healthcare professionals in making informed decisions regarding further diagnostic procedures and treatment plans. The classification model is applied at the patch level, with regions-of-interest extracted as patches from mammograms. The patches from lesions were extracted from both tightly fitting bounding boxes and larger bounding boxes surrounding the lesion, i.e. with different levels of zoom. Therefore, the influence of varying zoom levels of the lesion on the classifier's performance was examined.

Several interesting outcomes were observed in this project. Firstly, it was shown that the classifier's performance for the malignant class was inferior to that for the other classes. Moreover, it was also found that the tightness of the bounding box around the lesion has an impact on the classifier's performance. This motivates further study to improve the classifier's robustness against different accuracies of the annotation masks, that can be due to inter- and intra-observer variability.

To enhance the performance for malignant lesions, SinGAN-based data augmentation was employed to balance this underrepresented class. To the best of our knowledge, this project presents a novel investigation into the application of single-image generative models for breast cancer, representing the first-ever exploration of this approach in the field of breast imaging. Across all conducted experiments, the performance consistently improved when this data augmentation technique was used. However, other surprising outcomes were also observed. When using more SinGAN models to generate new data, it was expected to improve performance as more diversity was added to the training data. The performance did improve, but not in the expected trend. This outcome provides potential avenues for further analysis and research with the ultimate goal of achieving sufficient robustness for implementation in clinical practice.

This study introduces an innovative perspective on the utilisation of these models for breast cancer analysis, opening up new research possibilities and potential advancements in diagnostic methodologies. <sup>1</sup>

---

<sup>1</sup>The code of this project is available in the following link: [Master Project GitHub repository](#).



## Chapter 2

# Literature review

In recent years, the application of Convolutional Neural Networks (CNNs, LeCun, Bengio et al., 1995) has demonstrated exceptional performance in learning crucial features from mammography mass lesions for subsequent classification tasks (Abdelrahman et al., 2021). For instance, Arevalo et al., 2015 demonstrated remarkable results using CNNs for the classification of mammography mass lesions, utilising the same dataset as the one used in this study (BCDR dataset, Guevara Lopez et al., 2012). However, it is worth noting that their work exclusively focused on film mammograms. By incorporating both digital and film mammograms, this project approach aims to enhance the model's ability to generalise across different imaging modalities and improve its performance in real-world scenarios.

In a general overview, synthetic images generated by GANs Goodfellow et al., 2020 have demonstrated remarkable visual realism when applied to mammography. These synthetic images have shown potential in enhancing various subsequent tasks, such as cancer detection, tumor segmentation, and classification (Osuala et al., 2023a). Therefore, GANs can be studied and utilised for data augmentation or to address data imbalances, providing a means to generate additional training samples or correct the distribution of existing data. Utilising GANs in this context would enhance the robustness, generalisability, and performance of medical image analysis models, in applications such as nodule classification in mammography.

GANs (Goodfellow et al., 2020) are a class of Deep Learning models that consist of two neural networks: a generator and a discriminator. GANs is a framework for generating an estimate distribution based on an adversarial process between two models trained simultaneously. On the one hand, the generative model  $G$ , that captures the data distribution. On the other hand, a discriminative model  $D$ , captures the probability that a sample has been drawn from the training data rather than from the modelled distribution, so if it is a real sample or a generated ("fake") one. This simultaneous training corresponds to a minimax two-player game between  $G$  and  $D$ . The objective for  $G$  is to maximise the probability of  $D$  making a mistake, while  $D$  tries to classify fake and real images correctly.

In Goodfellow et al., 2020, they proposed an interesting analogy to understand the adversarial process, where the process is thought as a team of counterfeiters, analogous to the generator, trying to produce fake currency and use it without detection. The discriminative model would be analogous to the police, trying to detect the fake currency. This competition pushes both teams to improve their methods until the fake currency is indistinguishable from the original.

$G$  and  $D$  in the adversarial modelling framework are usually modelled as multilayer perceptrons. In this case, the system can be trained with backpropagation.

In order to learn the generator's distribution  $p_g$  over data  $\vec{x}$ , a prior distribution is defined on input noise variables:  $p_z(\vec{z})$ . A multilayer perceptron  $G$  represents a mapping from the noise variables to data space with the mapping  $G(\vec{z}; \Theta_g)$  with parameters  $\Theta_g$ . On the other hand, a second multilayer perceptron models the discriminator  $D$  as  $D(\vec{x}; \Theta_d)$  with parameters  $\Theta_d$  from the data space and outputs a probability.  $D$  is trained to maximise the probability of assigning the correct label to both training examples and samples from  $G$ .  $D(\vec{x})$  represents the probability that  $\vec{x}$  came from the actual data  $p_{data}$  rather than from the generated one  $p_g$ . A general scheme of GANs architecture is presented in Figure 2.1.  $G$  is simultaneously trained to minimise  $\log(1 - D(G(\vec{z})))$ .  $D$  and  $G$  play the following two-player minimax game with value function  $V(G, D)$ :

$$\min_G \max_D V(D, G) = \min_G \max_D \mathbb{E}_{\vec{x} \sim p_{data}(\vec{x})} [\log D(\vec{x})] + \mathbb{E}_{\vec{z} \sim p_z(\vec{z})} [\log(1 - D(G(\vec{z})))] \quad (2.1)$$

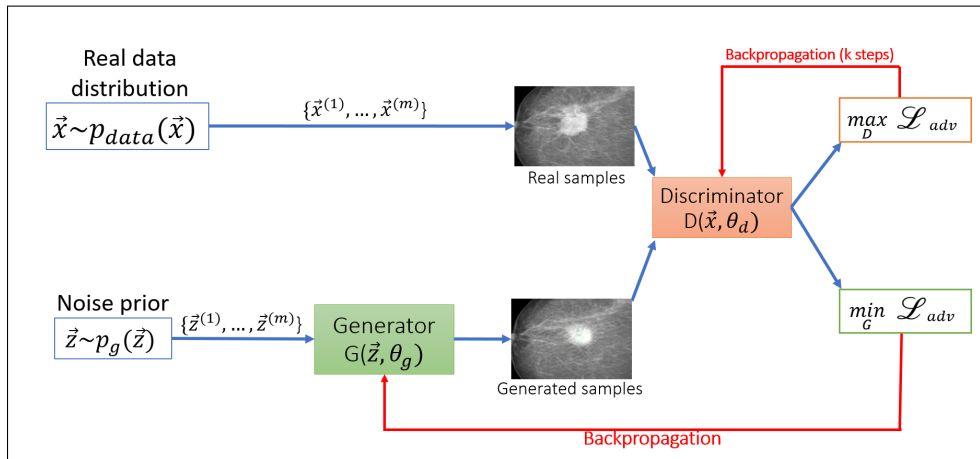


FIGURE 2.1: General pipeline of Generative Adversarial Networks (GANs). The generator network ( $G(\vec{z}; \Theta_g)$ ) takes in a random noise input and produces a synthetic output that is meant to be similar to the input data. The discriminator network ( $D(\vec{x}; \Theta_d)$ ) takes this generated data mixed with the real samples and it learns to classify it as either "real" or "fake" data. Both networks are trained simultaneously through backpropagation during a number of iterations in an adversarial process.

However, training GANs typically requires large datasets, which can be time-consuming to collect. Moreover, in certain domains such as healthcare, acquiring large annotated datasets is often impractical due to the need for expert annotations, which are both costly and time-consuming. This challenge is particularly relevant in the medical domain. To address these limitations, the SinGAN framework (Shaham, Dekel and Michaeli, 2019) offers a unique approach. Unlike traditional GANs, which are trained on large datasets, SinGAN is designed to be trained on a single image. This substantially reduces the data requirements and computational burden associated with training.

There have been several successful studies in breast cancer utilizing GAN architectures. In their paper, Lee and Nishikawa, 2022 demonstrated the feasibility of employing mammograms generated by Conditional GANs to detect mammographically-occult (MO) cancer in women with dense breasts. Additionally, Eric et al., 2018 augmented the dataset with high-resolution synthetic mammogram patches generated by a class-conditional GAN. They illustrated that a ResNet-50 classifier, trained with GAN-augmented training data, achieved a higher AUC compared to the same model trained solely on traditionally augmented data. Furthermore, Wu, Wu and Lotter, 2020 investigated the augmentation of the original training set with GAN-generated samples, resulting in a significant improvement in malignancy classification performance on a test set of real mammogram patches. Once demonstrated that synthetic data from generative models can enhance the performance of data-hungry Deep Learning models in medical imaging. To facilitate this process, Osuala et al., 2023b developed medigan, an open-source library of pretrained generative models for medical image synthesis. medigan enables researchers and developers to easily create, augment, and adapt their training data in medical imaging.

On the other hand, the SinGAN model has seldomly been used in medical imaging and, to the best of our knowledge, this is the first study done for breast cancer diagnosis using this method. Remarkably, Thambawita et al., 2022 showed promising results of SinGAN framework for medical applications. The authors successfully applied the SinGAN framework to cancer imaging for polyp segmentation. Their work not only involved generating synthetic images but also generating corresponding masks as an additional channel to the image. This innovative application highlighted the potential of utilising SinGANs in the medical context.





## Chapter 3

# Methods and materials

### 3.1 BCDR dataset

The Breast Cancer Digital Repository (BCDR, Guevara Lopez et al., 2012) is an extensive accessible repository that comprises annotated cases of breast cancer patients from the northern region of Portugal. BCDR provides both normal and annotated patient cases, including mammography lesion outlines, anomalies observed by radiologists, and relevant clinical data. Most mammograms with suspicious lesions contain an annotated mask indicating the affected region, created by a radiologist.

This study is focused on the classification of lesions in patches extracted from digital and scanned film mammograms. The objective is to enhance the robustness and generalisability of the classification model by incorporating data from both sources. Specifically, the research looks for classifying patches into: healthy (no lesion), malignant, and benign. The classification into malignant and benign classes is based on biopsy results, offering a comprehensive assessment of the nature of the lesions.

### 3.2 Mammogram patch extraction

#### 3.2.1 Rationale

The extraction of patches serves two primary purposes in this study. Firstly, the patches are extracted to facilitate the implementation of a sliding window procedure for detection. This approach enables the model to evaluate multiple regions within an image and make predictions based on individual patches. Secondly, the extraction of patches aims to predict, based on a quick annotation mask of the lesion's location, whether the lesion is more likely to be benign or malignant even before a biopsy is performed. This pre-biopsy prediction can provide valuable information to assist healthcare professionals in making informed decisions regarding further diagnostic procedures or treatment plans.

Additionally, the patches containing lesions are extracted with varying percentages of adjacent healthy tissue. This approach serves two purposes. Firstly, it allows for the exploration of the model's robustness against different window sizes used in the sliding window procedure for lesion detection. By including patches with varying amounts of healthy tissue surrounding the lesion, the model can adapt to different window sizes and improve its detection capabilities.

Secondly, it aims to analyse the model's adaptability to imperfect annotations, enabling it to make accurate classifications despite variations in the quality or accuracy of the annotation masks. This includes considering both inter-observer and

intra-observer variability presented by Elmore et al., 1994. Inter-observer variability refers to the robustness of the model against different annotations provided by different experts. It ensures that the model can consistently produce accurate results regardless of the variations in annotations among different experts. Similarly, the model should also account for intra-observer variability, which involves assessing its performance when the same expert provides annotations on different days. This variability may arise due to factors such as varying expertise levels, changing subjective judgment, or inconsistencies in the annotation process.

By including patches with different amounts of adjacent healthy tissue, the model becomes more robust and can handle variations in the precision or consistency of the annotation masks. This approach helps to evaluate the model's performance dependency on precise annotations and its ability to generalise well in real-world scenarios where annotations may be less precise or consistent. In doing so, the model can effectively handle imperfect annotations and maintain accurate classifications, ensuring its reliability in practical applications.

### 3.2.2 Technique

In this section, the complete procedure for extracting patches from the mammograms is explained. The BCDR repository Guevara Lopez et al., 2012 provides a folder of normal (healthy) mammograms exclusively for the digital dataset. Thus, healthy patches from digital mammograms are extracted from breasts without any annotation indicating the presence of a lesion. Bounding boxes were generated using a sliding window approach within completely healthy breast images, ensuring that these patches never contain more than 50% of background pixels. It was configured with a zero stride, so adjacent patches do not overlap. In Figure 3.1, an example of a digital normal mammogram is included, with the boxes used in this case for extracting healthy patches.

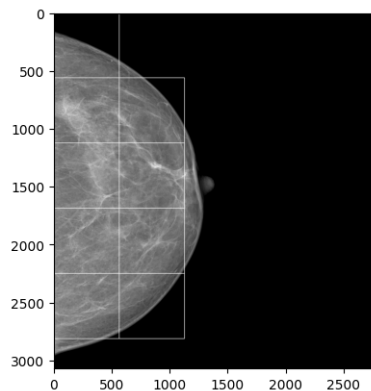


FIGURE 3.1: Example of a normal digital mammography image illustrating the procedure for extracting healthy digital patches.

In contrast to the digital subdataset, the film subdataset does not have a separate folder for normal mammograms. Therefore, to extract healthy patches from film mammograms, patches of healthy tissue from suspicious film images were selected. To achieve this, a margin of at least 20 pixels was ensured between the extracted healthy patch and the bounding box of the lesion, while also satisfying the constraint of not containing more than 50% background pixels. Since film images are more

limited in quantity, a stride of 65 pixels was utilised to obtain healthy patches. In Figure 3.2, an example of a film suspicious mammogram is included, with the boxes used in this case for extracting healthy patches.

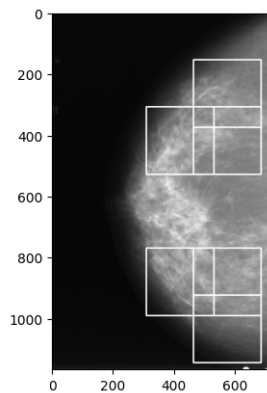


FIGURE 3.2: Example of a suspicious film mammography image illustrating the procedure for extracting healthy film patches.

Non-healthy patches are crops that contain the Region of Interest (ROI) of both malignant and benign lesions of any type present in the datasets, including masses, calcifications, microcalcifications, and architectural distortions. The annotated lesion masks were used, created by a radiologist, to create bounding boxes that enclose them, from which square patches were extracted. If the margin extends beyond the border of the mammogram, a translation is performed to ensure the patch remains fully within the limits of the mammogram.

Figure 3.3 displays a sample of a suspicious film mammogram along with its corresponding annotation mask of the lesion. This is the same sample introduced in Figure 3.2 for healthy patch extraction. Additionally, Figure 3.4 showcases a sample of a digital mammogram with the corresponding annotation of the lesion ROI.

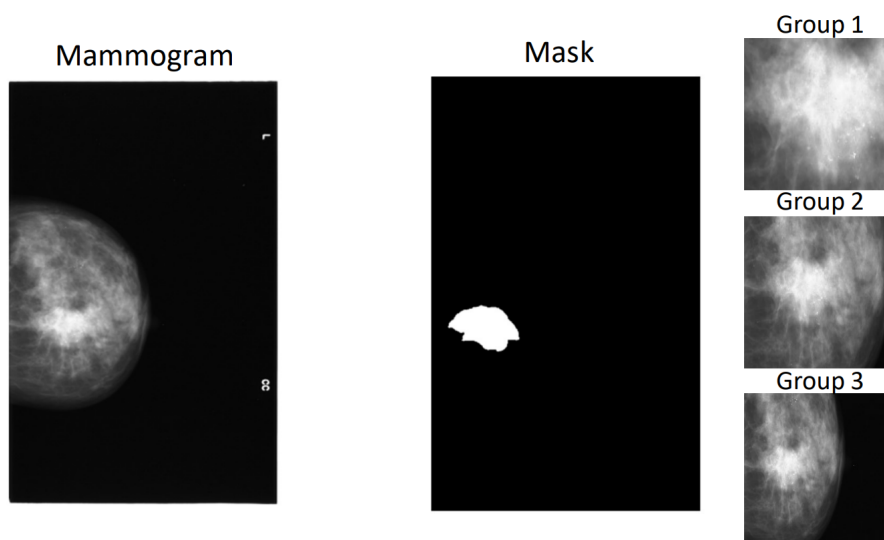


FIGURE 3.3: Film mammogram and the corresponding annotation mask. In the third column, the patches extracted from them from the three different zoom levels, being the Group 1 the most accurate ROI bounding box.

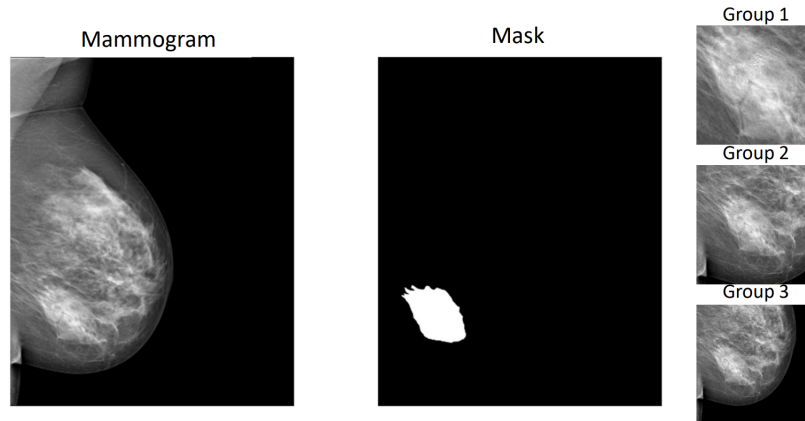


FIGURE 3.4: Digital mammogram and the corresponding annotation mask. In the third column, the patches extracted from them from the three different zoom levels, being the Group 1 the most accurate ROI bounding box.

Each lesion has three patches with different levels of zoom, capturing varying percentages of adjacent healthy tissue for the purposes explained in Section 3.1. Group 1 patches correspond to the original bounding box defined around the annotated mask. Group 2 and 3 capture patches with double and triple the height and width of the original bounding box, respectively. While group 1 patches capture the minimum subjacent healthy tissue, group 3 patches capture more subjacent healthy tissue. Figure 3.5 shows sample patches of lesions extracted from different zoom levels of both digital and film mammograms.

In the final step of patch extraction, they were resized to 224x224 pixels using interpolation via the *OpenCV: Open Source Computer Vision Library* n.d. Each generated patch has a unique ID. The metadata collected for each patch includes the unique patient ID, mammogram format (digital or film), zoom group (1, 2, or 3, and 0 for healthy patches), image view type, breast density, biopsy result, and boolean variables indicating the presence of microcalcifications, calcifications or nodules. The BCDR dataset includes samples from different image view types: RCC (Right Cranio-Caudal), LCC (Left Cranio-Caudal), RMLO (Right Medio-Lateral Oblique), and LMLO (Left Medio-Lateral Oblique). Breast density is determined based on *Breast Imaging Reporting & Data System (BI-RADS®) 2023*, which classifies it as A, B, C, or D. The BCDR dataset utilises the previous version of BI-RADS, in which the numbers 1, 2, 3, and 4 corresponded to the letters A, B, C, and D respectively.

For more details, you can find the complete code at the following link: [Master Project GitHub repository](#).

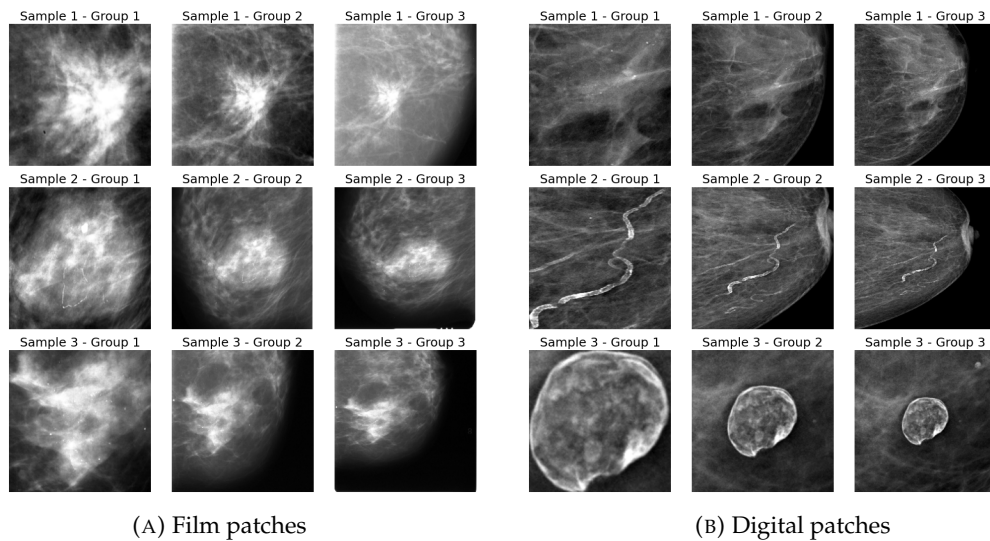


FIGURE 3.5: On the left, three samples of lesions extracted from scanned film mammograms at different zoom scales (Group 1, Group 2, and Group 3). All three lesions are benign: the first one is a nodule, the second one is a calcification with a nodule, and the third one is another distortion. On the right, three samples of lesions extracted from digital mammograms, also at different zoom scales. The first sample is a malignant nodule, while the second and third samples are benign calcifications.

### 3.3 Analysis of the data

In this subsection, the metadata distributions of the patch dataset that was created are analysed. The dataset comprises a total of 5408 patches from 473 patients. The age distribution of these patients is visualised in Figure 3.6, where it can be observed that the highest density of patients falls within the age range of 50 to 65 years old. This concentration of patients in the middle-aged group aligns with the fact that the risk of developing breast cancer increases significantly for women over the age of 40 (*WHO: Breast Cancer Fact Sheet 2021*).

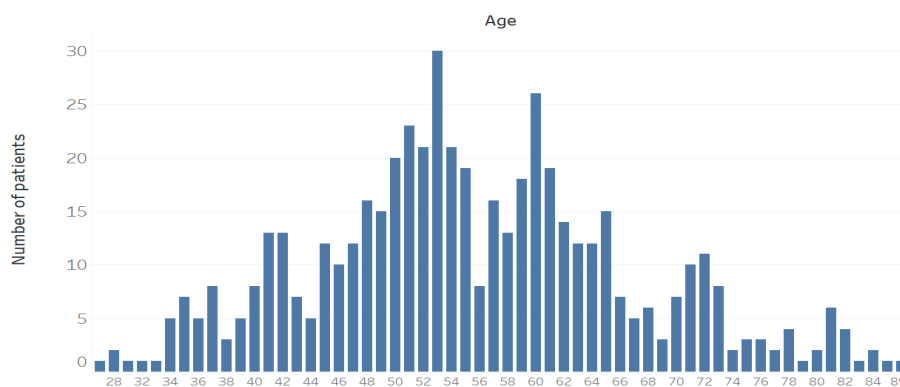


FIGURE 3.6: Distribution of ages of the different patients included in the patch dataset generated from the BCDR dataset. There are 473 in total.

The research is centered on classifying lesions into three distinct categories: healthy (no lesion), malignant, and benign. The classification into malignant and benign

classes is based on biopsy results. The distribution of the three classes is visualised in Figure 3.7a. The majority of patches in the dataset are categorised as normal, indicating the absence of suspicious masses. However, among the suspicious patches, they are further divided into the benign and malignant classes. It is evident that the malignant class is the minority, presenting an imbalanced distribution that will receive special attention in this project. According to Figure 3.7b, there are a total of 984 lesions, and for each of these lesions, three patches are available, corresponding to the three groups of zoom scale.

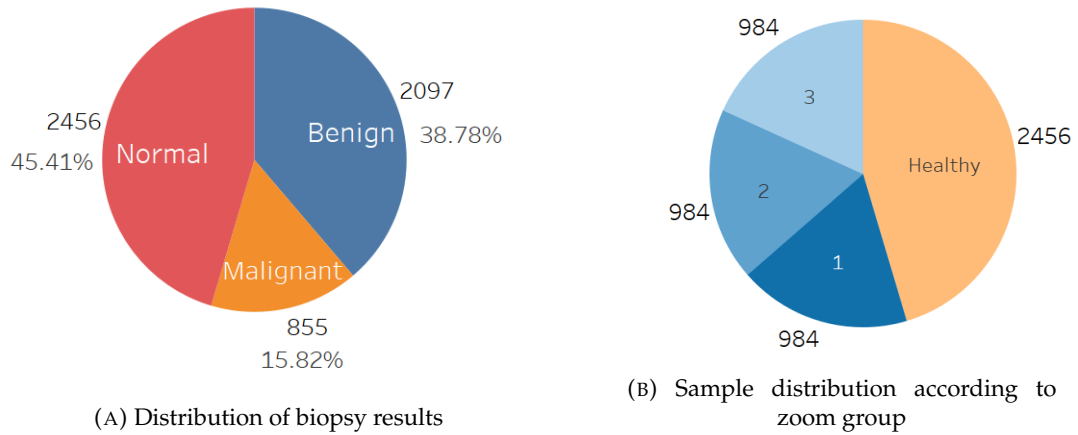


FIGURE 3.7: On the left image, the distribution of patches based on biopsy classification. The majority of patches are normal with no suspicious mass, while the suspicious patches are divided between benign and malignant. On the right, the distribution of patches based on the zoom group of the lesions, along with the total number of patches in each group.

According to Nalawade, 2009, a biopsy can be avoided if the calcifications appear absolutely benign on mammography and the patient can be followed-up with annual screening mammography. Although the presence of calcification in a lesion patch was not taken into consideration for the classification pipeline, this feature has been explored for subsequent analysis. In the bar chart shown in Figure 3.8, there are four different classes for the lesions. The lesions can be classified as nodules, calcifications, nodules with calcifications, or other anomalies (microcalcification, axillary adenopathy, architectural distortion, or stroma distortion). Based on this Figure, 54.51% of the benign samples exhibit calcification without nodules, while only 7.02% of the malignant samples demonstrate the same characteristic. Additionally, the majority of malignant lesions (55.44%) are nodules. Considering this data from a frequentist approach, it can be inferred that a given calcification sample is more likely to be benign.

Regarding the breast density, as shown in Figure 3.9a, it can be observed that the density groups are unbalanced, with groups 1 and 4 being the minority categories. According to *Canadian Cancer Society 2023*, approximately 40% of women fall into density category 2, another 40% in category 3, 10% in category 1, and 10% in category 4, which aligns reasonably well with the distribution observed in the patch dataset. Regarding the type of image view, as depicted in Figure 3.9b, there is no significant imbalance among the different views.

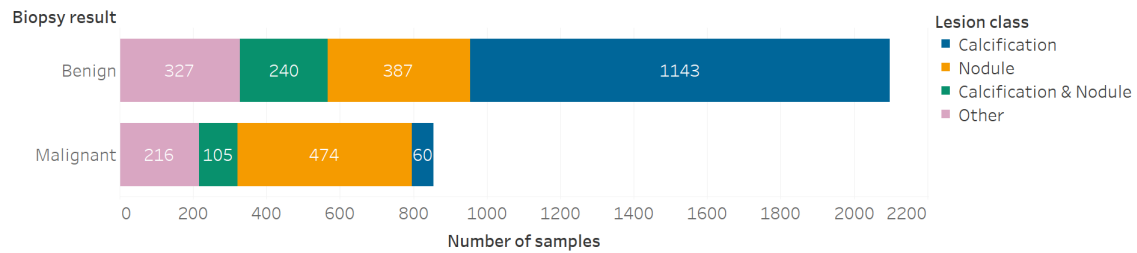
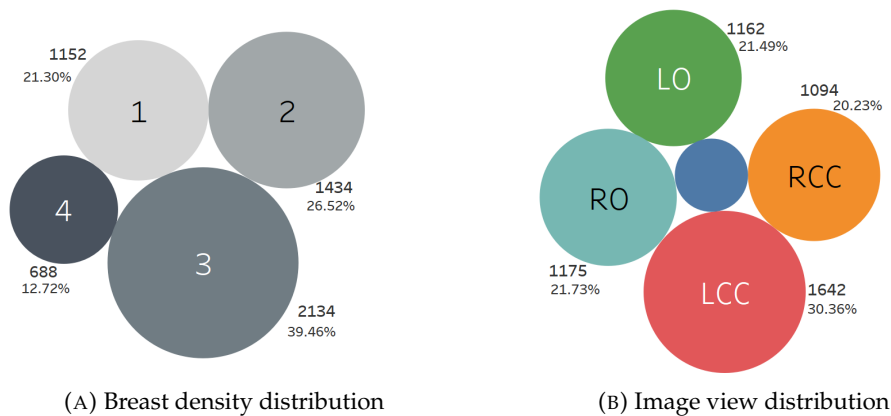


FIGURE 3.8: Distribution of different classes for lesions with calcification, with nodule, with both calcification and nodule or and without calcification for each classification type (benign and malignant).



(A) Breast density distribution

(B) Image view distribution

FIGURE 3.9: On the left image, the distribution of breast density according to *Breast Imaging Reporting & Data System (BI-RADS®) 2023*, with 1 indicating lower density and 4 indicating higher density. On the right, the distribution of mammography view types: RCC (Right Cranio-Caudal), LCC (Left Cranio-Caudal), RMLO (Right Medio-Lateral Oblique), LMLO (Left Medio-Lateral Oblique) and some without specification (blue area).

Both film and digital mammograms are being utilised in this research. However, as illustrated in Figure 3.10, there is a larger proportion of patches derived from digital images. Among the healthy patches, approximately 41.5% are from film mammograms, while for the patches with a suspicious lesion, around 39.8% originate from film mammograms. Thus, there is a comparable level of imbalance observed for both healthy and lesion patches. It is worth noting that this imbalance is not substantial, as the film class constitutes approximately 40% of the total patch distribution.



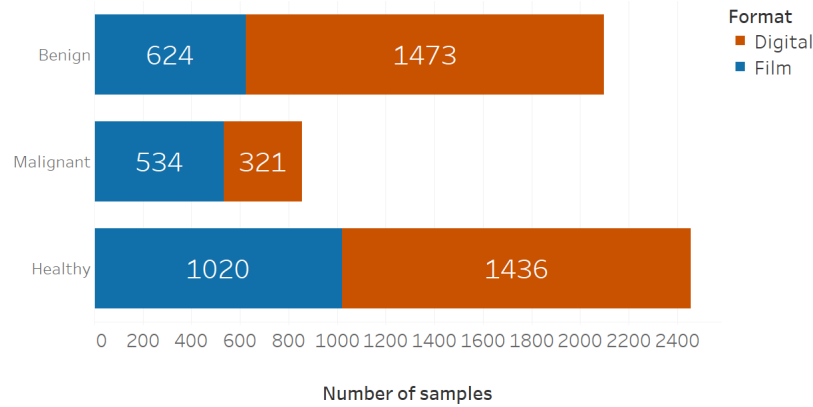


FIGURE 3.10: Distribution of the format (digital or film) of the mam-mography for healthy and suspicious patches.

### 3.4 Data augmentation with SinGAN

As depicted in Figure 3.7a, the malignant biopsy result represents a minority class within the dataset. As I will elaborate in Section 4, this class imbalance poses a challenge and adversely affects the classification performance for malignant samples. Given the objective of achieving a robust pre-biopsy classification outcome, The potential of augmenting the malignant samples using synthetic images generated from a generative model was explored. This approach aimed to bolster the representation of the malign class and improve the classification results.

For this purpose, SinGAN (Shaham, Dekel and Michaeli, 2019) was used, a promising framework that addresses the challenge of data scarcity in cancer imaging. SinGAN provides a solution by generating multiple synthetic images from a single training image. This technique enables the augmentation of the dataset, enhancing the representation and diversity of the malignant class for improved classification performance.

#### 3.4.1 SinGAN model architecture

The architecture of SinGAN reframes the traditional idea of GANs. Instead of modeling the distribution of a set of images, as initially proposed by Goodfellow et al., 2020, SinGAN focuses on modeling the distribution of different overlapping patches within a single image, across different scales. This hierarchical approach enables SinGAN to capture fine-grained details while preserving global structures.

They proposed a pyramidal structure with one GAN (one discriminator and one generator) per each of the  $N$  scales. Every generator takes a noisy image as an input and produces a new image of the same size as an output. The output of the generator  $n$  is then the input of the discriminator on scale  $n$ , just like in a classical GAN. The output of the scale  $n$  is resized to a higher resolution and – with some additional noise – added to the generator of scale  $n-1$ . In this way, the  $N$ -th scale defines the overall structure of the image with the coarsest resolution and the first scale defines the finest structures and details.

For instance, let's consider training SinGAN on a natural image of a city skyline. In this scenario, the  $N$ -th scale could define the positioning and arrangement of



buildings in the generated image. The  $(\mathcal{N}-1)$ -th scale would determine the overall shape and structure of each building, while the  $(\mathcal{N}-2)$ -th scale could contribute to the variations in architectural styles and features. Additionally, the  $(\mathcal{N}-3)$ -th scale might influence the distribution of windows and doors, while the  $(\mathcal{N}-4)$ -th scale could control the placement and density of people or vehicles within the cityscape.

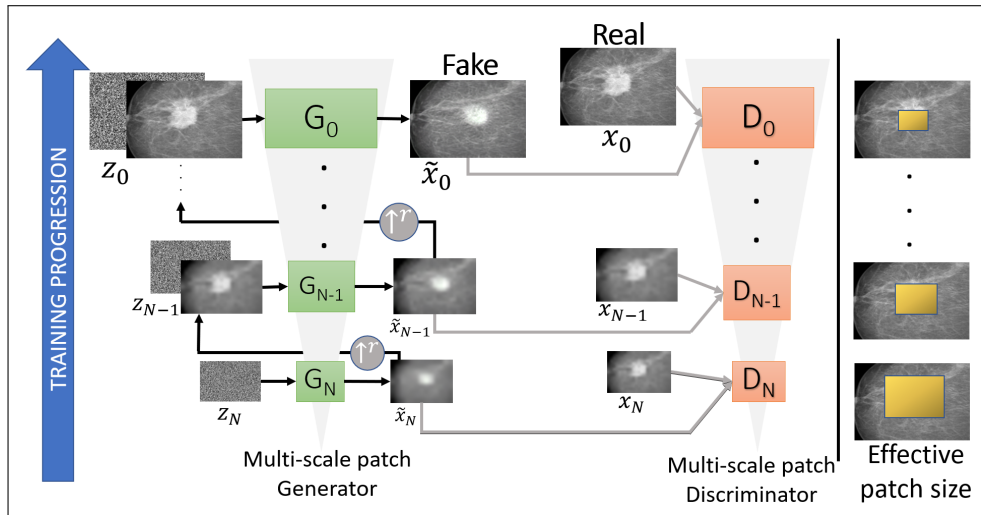


FIGURE 3.11: General pipeline of the SinGAN framework. Figure adapted from Shaham, Dekel and Michaeli, 2019.  $\mathcal{N}$  GANs are operating on different scales. The training process starts with the coarsest scale and progresses to the finest scale. Each GAN in the hierarchy learns to generate realistic images at its respective scale, capturing both global and local details. At each scale  $n$ , the image from the previous scale,  $\tilde{x}_{n+1}$ , is upsampled and added to the input noise map,  $z_n$ . The result is fed into the generator ( $G_n$ ), whose output is a residual image that is added back to ( $\tilde{x}_{n+1}$ )  $\uparrow^+$ . This is the output  $\tilde{x}_n$  of  $G_n$ .

The training procedure of the SinGAN structure addresses the challenges of overfitting that arise from training on a single image through the simplicity of the generator and discriminator architectures. The generator consists of only a few convolutional layers, which limits its effective receptive field, ensuring variance at each scale and preventing the generator from capturing the global structure of the image. As the image is upsampled at each step, the relative effective receptive field decreases, shifting the focus from the global structure at the coarsest scale to the fine textures at the finest scale. Similarly, the complexity of the discriminator is restricted to prevent it from memorising the real image.

Another crucial aspect is the mapping of the noisy image. To ensure that the generator functions as an identity function when no noise is added, a modified loss function is employed. In addition to the adversarial loss, which penalises deviations from the distribution of real patches, a reconstruction loss term is added. This term penalises deviations from the identity function when the input noise is set to zero.

The training of the hierarchical GAN structure is performed sequentially, starting from the coarsest scale  $\mathcal{N}$  and progressing to the finest scale 1. The generator at each scale is trained against its corresponding discriminator and reconstruction task. Once the training of one scale converges, the next finer scale is trained. Therefore,

instead of training a single GAN,  $\mathcal{N}$  GANs are trained sequentially, which is a trade-off for training on a single sample.

The training loss for each GAN incorporates two terms (3.1): the adversarial loss ( $\mathcal{L}_{adv}$ ) and the reconstruction loss ( $\mathcal{L}_{rec}$ ).

$$\min_{G_n} \max_{D_n} \mathcal{L}_{adv}(G_n, D_n) + \alpha \mathcal{L}_{rec}(G_n) \quad (3.1)$$

The adversarial loss penalises the difference between the distribution of patches in the real image  $x_n$  and the distribution of patches in the generated samples  $\tilde{x}_n$ . This loss term helps the generator to produce samples that resemble the real data distribution. In the paper by Shaham, Dekel and Michaeli, 2019, the Wasserstein GAN with Gradient Penalty (WGAN-GP) loss function, proposed by Gulrajani et al., 2017, is used for training the generator and discriminator. The authors observed that this loss function, compared to other GAN variants, enhances training stability. The formula of the WGAN-GP loss function is:

$$\mathcal{L} = \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{g}}} [f(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r} [f(\mathbf{x})] + \lambda \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{g}}} [(\|\Delta_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}})\|_2 - 1)^2] \quad (3.2)$$

In Equation 3.4.1, the first two terms are the original WGAN loss and the right term is the gradient penalty with  $\lambda$  the penalty coefficient.

When there is no noise added to the image, each generator should just be the identity function, returning the input image  $x_n$ . This is achieved by the reconstruction loss, which penalises deviations from the identity function, given zero noise. If the noise images added in each of the  $n$  scales are chosen so that  $\{z_{\mathcal{N}}^{rec}, z_{\mathcal{N}-1}^{rec}, \dots, z_0^{rec}\} = \{z^*, 0, \dots, 0\}$ , where  $z^*$  is some fixed noise map. Denoting by  $\tilde{x}_n^{rec}$  the generated image at the  $n$ th scale when using these noise maps. Then for  $n < \mathcal{N}$ :

$$\mathcal{L}_{rec} = \|G_n(0, (\tilde{x}_{n+1}^{rec})) \uparrow^r - x_n\|^2$$

and for  $n = \mathcal{N}$ :  $\mathcal{L}_{rec} = \|G_{\mathcal{N}}(z^*) - x_{\mathcal{N}}\|^2$ .

In summary, SinGAN is a stacked ensemble of  $\mathcal{N}$  GANs that operate on different scales of the same image. Despite the requirement of only a single image as training input, the training process involves sequentially training  $\mathcal{N}$  GANs, which is the trade-off for leveraging the capabilities of SinGAN.

### 3.4.2 SinGAN implementation

The synthetic dataset was generated using four different SinGAN models trained on distinct images. To enhance robustness against the digital and film formats, two digital and two film patches from malignant lesions for training each of the four SinGAN models were employed. This set of four SinGAN models will be referred to as 'Set A' and was exclusively generated from patches belonging to zoom group 3, which contains a higher percentage of adjacent healthy tissue. In addition, another synthetic dataset was generated by training four different SinGAN models (referred to as 'Set B') on four different training images, also from zoom group 3 and two film and two digital. The purpose of this was to repeat the experiment using a different set of SinGAN models and assess the consistency of the results.

Furthermore, for a different experiment, four different SinGAN models (referred to as 'Set C') were trained on four single patches from lesions (two film and two digital), specifically from zoom group 1, which includes accurately annotated bounding boxes around the lesions. The training images used for this experiment are presented in Figure 3.12, organised by format. The first and second rows correspond to Set A and Set B, respectively, both derived from zoom group 3. The third row corresponds to Set C, which is associated with zoom group 1.

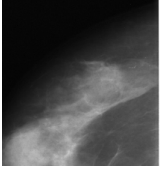
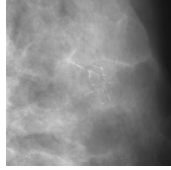
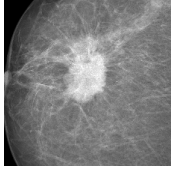
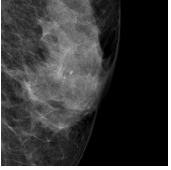
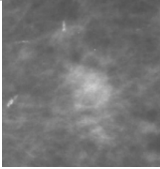
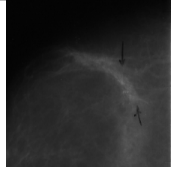
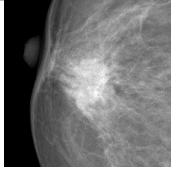
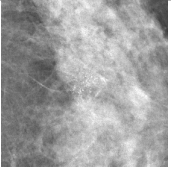

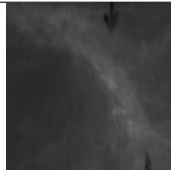
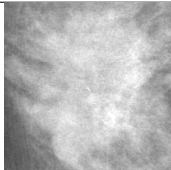
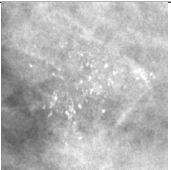
SET	FILM FORMAT		DIGITAL FORMAT	
<b>A</b>				
<b>ID</b>	51	168	2180	3884
<b>B</b>				
<b>ID</b>	13	77	2789	3941
<b>C</b>				
<b>ID</b>	11	75	2787	3939

FIGURE 3.12: Original patches from malign lesions used for training each SinGAN model. The first two rows consist of patches from the zoom level group 3, which has the highest percentage of adjacent healthy tissue. The last row contains a set of 4 patches from the zoom level group 1, which have tight bounding boxes around the lesions. Each set of 4 patches is labeled as Set A, Set B, and Set C, respectively. Within each set, there are 2 patches extracted from film mammograms and 2 from digital mammograms. The ID corresponds to a unique label assigned to each patch extracted from the BCDR dataset.

In this study, the public repository provided by the original SinGAN paper *Official pytorch implementation of the paper: "SinGAN: Learning a Generative Model from a Single Natural Image" 2019* was used, with adaptations made for some of the hyperparameters. A pyramid scale factor of 0.8 was chosen, this means that the resolution of the image when passing to the next scale is reduced by a 20%. Therefore, the pyramid had 11 scales and 30 epochs were computed for each scale. The default values were used for the rest of hyperparameters. By increasing the number of scales, finer details in the generated images can be captured. Similarly, a higher number of epochs allows for more training iterations, potentially leading to improved convergence and overall image quality. It is important to note that a trade-off was considered between computational cost and image quality during parameter selection. Some samples of the generated images with the SinGAN models trained are presented in Table 3.13.

The training process for each SinGAN model took approximately 3.5 hours, using the computational resources of the Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM) server with the following service available: NVIDIA RTX 2080 Super 8GB GPU. This enabled efficient training and generation of synthetic images within a reasonable time frame.

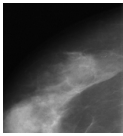
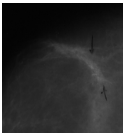
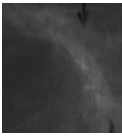
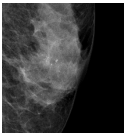
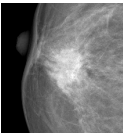
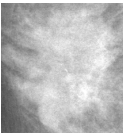
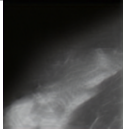
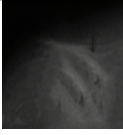
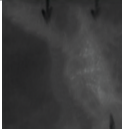
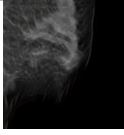
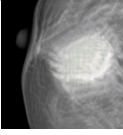
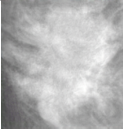
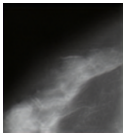
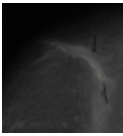
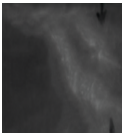
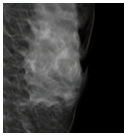
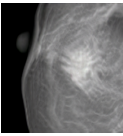
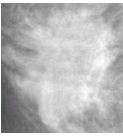
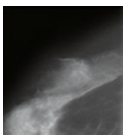
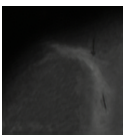
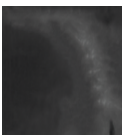
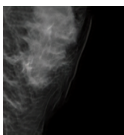
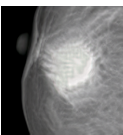
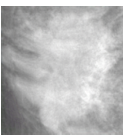
ORIGINAL						
ID	51	77	75	3884	2789	2787
SinGAN						
						
						

FIGURE 3.13: Some samples of generated images with different SinGAN models. The corresponding image used for training the model is presented in the first row. The ID corresponds to a unique label assigned to each patch extracted from the BCDR dataset.

### 3.4.3 SiFID as evaluation metric

The Fréchet Inception Distance (FID) (Heusel et al., 2017) metric is commonly used to evaluate the quality of generated images in generative models. It is a metric that calculates the Fréchet distance between feature vectors calculated for real and generated images. It indicates how similar the two groups are in terms of statistics on computer vision features of the raw images calculated using an Inception model v3 trained on the ImageNet used for image classification. This distance metric is formulated as follows:

$$d^2((\mu_1, C_1), (\mu_2, C_2)) = \|\mu_1 - \mu_2\|_2^2 + \text{Tr}(C_1 + C_2 - 2(C_1 C_2)^{1/2}) \quad (3.3)$$

The parameters  $\mu_1$  and  $\mu_2$  refer to the feature-wise mean of the real and generated images and parameters  $C_1$  and  $C_2$ , to the covariance matrices for real and generated feature vectors.

Lower scores indicate the two groups of images are more similar, or have more similar statistics, with a perfect score being 0.0 indicating that the two groups of images are identical.

In this study, the Fréchet Inception Distance (FID) metric was not used. Instead, SiFID, a metric based on the FID metric, was used as a substitute. SiFID stands for

Single Image FID and was proposed by Shaham, Dekel and Michaeli, 2019. Instead of using the activation vector after the last pooling layer in the Inception model (a single vector per image), SiFID uses the internal distribution of deep features at the output of the convolutional layer just before the second pooling layer (one vector per location in the map). The SiFID is the FID between the statistics of those features in the real image and in the generated sample.

### 3.5 Classification pipeline

The research is based on a classification pipeline for distinguishing patches from mammograms into three distinct categories: healthy (no lesion), malignant, and benign. The classification of lesions into malignant and benign classes is based on biopsy results, providing a reliable ground truth for training the model. To accomplish this task, a pre-trained ResNet50 model that was originally trained on the ImageNet dataset from *PyTorch: models and pre-trained weights 2023* was utilised. By transfer learning, I could leverage the knowledge learned from the ImageNet dataset to improve the classification performance on my specific task.

The choice of using a Residual Network (ResNet) (He et al., 2016) was motivated by its demonstrated better performance over traditional deep neural networks. Deep neural networks often encounter the vanishing gradient problem, where gradients diminish significantly as the network becomes deeper, hindering weight updates and impeding further training progress. Residual Networks effectively address this issue by incorporating skip connections, also known as "shortcuts," between every two layers. These direct connections allow the activation from one layer to be fed directly to another layer, facilitating the flow of information and preserving the learning parameters in deeper layers. ResNet50 specifically refers to a Residual Network architecture comprising 50 layers, including convolutional layers, pooling layers, fully connected layers, and the aforementioned residual connections. In Appendix 6.1, a diagram of the complete model is included.

The ResNet50 model was initialised with the default weights and then added a last linear layer, which adapts the output to match the number of classes in the problem addressed. This number of classes were three (healthy, malignant, and benign) for the multiclass problem and two, for the binary one. To optimise the training process and focus on fine-tuning the last layer for improved classification performance, only the parameters of the last layer were kept trainable. For the multiclass task, there were finally 6147 trainable parameters.

To fine-tune the model's hyperparameters, an initial binary pipeline was adopted to classify samples as either healthy or suspicious. The benign and malignant classes were combined into a single category for this purpose. After fine-tuning, the following hyperparameters were fixed to ensure consistent comparisons across experiments: a batch size of 128, the adaptive moment (Adam) optimiser with default beta parameters ( $\beta_1=0.9$  and  $\beta_2=0.999$ ), and a learning rate scheduler that progressively decayed the learning rate. The scheduler had a step of 5 epochs and a gamma value of 0.1, which represents the factor by which the learning rate is multiplied in each step. For the binary classification problem, the Binary Cross Entropy loss function was utilised, which expression is:

$$\text{BinaryCE}(p) = -y \cdot \log(p) + (1 - y) \cdot \log(1 - p), \quad (3.4)$$

where  $y$  is the actual label (1 or 0) and  $p(y)$  is the predicted probability of the positive class. For the multiclass problem, the same hyperparameters were used, but with the Categorical Cross Entropy loss function as the criterion, which is similar to the Binary Cross Entropy loss, but adapted for more classes:

$$\text{CategoricalCE}(p) = - \sum_{i=1}^N (y_i \cdot \log(p_i)), \quad (3.5)$$

where  $y_i$  represents the ground truth of class  $i$  and  $p_i$  represents the predicted probability of class  $i$ . Each experiment ran for 100 epochs and the model with the best validation loss was selected. They were evaluated using a train-validation-test split across three folds, ensuring that each patient was present in only one of the sets.

### 3.5.1 Evaluation of the classification

In order to evaluate the performance of the classifier, various metrics were selected. Firstly, the accuracy metric was considered. It was chosen as it represents the ratio of correct predictions to the total number of input samples:

$$\text{Accuracy} = \frac{\# \text{ of correct predictions}}{\text{Total number of predictions}}$$

By using this metric, an overall idea of the classifier's performance can be obtained. However, in cases where there is a significant class imbalance, a high accuracy score can misleadingly imply good performance. This issue becomes more critical when the cost of misclassifying the minor class samples is substantial. In the dataset utilised in this study, the malignant class constituted the minority, and the cost associated with misclassification, leading to underdiagnosis, was considerable. Consequently, a different metric was also employed to provide a more comprehensive evaluation of the classifier's performance.

To gain a better understanding of the performance, the ROC curve and ROC AUC score were computed for each experiment. The ROC (Receiver Operating Characteristic) curve is a graphical representation that illustrates the classification model's performance at various classification thresholds by plotting the True Positive Rate (proportion of actual positives correctly identified) against the False Positive Rate (proportion of actual negatives incorrectly identified as positives). AUC measures the complete two-dimensional area under the complete ROC curve. Therefore, it is an integral calculation from (0,0) to (1,1) of the curve. Its expression is as follows:

$$\text{AUC} = \int_{(0,0)}^{(1,1)} \text{TPR}(\text{FPR}) \delta(\text{FPR}) \quad (3.6)$$

The selection of these metrics for comparing the performance of the different experiments was motivated by their ability to demonstrate the separability of the classes across all possible thresholds, thereby indicating how effectively the model classifies each class. Other metrics that consider the outcomes as discrete may not provide as comprehensive insights. For the multiclass problem, the ROC curve was adapted using the One versus Rest (OvR) strategy. Under the OvR strategy, one class is treated as the "positive" class while considering all other classes as the "negative" class. This transformation reduces the multiclass problem to a series of binary ones,



so this process is repeated for each class present in the dataset. Consequently, for a dataset with three classes, three different OvR scores, along with their respective ROC curves, are obtained. In the experiments conducted in Section 4, three AUC scores and three ROC curves were computed, each corresponding to a specific class.

### 3.5.2 Malignant class balancing

The performances of two different approaches in addressing the class imbalance issue was compared and their effectiveness in improving the classification results was determined. The first approach involves data augmentation using synthetic images generated by SinGAN, while the second approach utilises sample weights.

For the experiments involving synthetic data, the synthetic samples were exclusively introduced into the training set. It is important to note that special care was taken to maintain consistency between the labels and metadata of the synthetic data and the corresponding original images used to train the SinGAN model. This ensured that the synthetic data accurately reflected the characteristics and annotations of the original images. Additionally, thorough attention was given to ensuring that the samples used to train the SinGAN and generate the synthetic data were exclusively present in the training set. This preserves the independence and reliability of the results. Finally, all data, both synthetic and original, underwent the necessary preprocessing required by the pretrained model.

For experiments that do not involve the use of synthetic samples to balance the malignant class, the Weighted Random Sampler ([PyTorch: Weighted Random Sampler 2023](#)) technique was applied to both the train and validation sets. The Weighted Random Sampler is a method that addresses class imbalance by assigning weights to each sample during the sampling process, ensuring that less frequent classes receive higher probabilities of being selected. By using this technique, it was aimed to mitigate the impact of class imbalance on the training and validation stages.

In cases where synthetic samples were introduced to balance the classes within the training set, the Weighted Random Sampler was exclusively utilised for the validation set. This decision was made to preserve the integrity of the synthetic samples and avoid introducing biases during the validation process. The performances of these two approaches in addressing the class imbalance issue will be compared and their effectiveness in improving the classification results will be determined.

### 3.5.3 Varying the zoom level group with which the classifier is trained and tested

The impact of the accuracy of lesion annotation on the performance of the classifier is explored through the conducted experiments. The area of adjacent tissue included in the patch as a lesion is reduced as the annotation accuracy increases. For this purpose, different combinations of zoom level group sets are used to train and test the classifier. The training set remains the same for validation in all cases.

For example, in one of the experiments, the classifier is trained using the set of lesions from zoom group 3 (with more healthy adjacent tissue) and subsequently tested with zoom group 1 (containing less healthy adjacent tissue). This allows us to assess how well the classifier is able to recognise the patterns of the lesion itself with

minimal adjacent tissue (group 1) when it has been trained with a greater amount of healthy tissue (group 3).

In this experiment, it was also investigated how data augmentation using SinGAN-generated images can enhance the robustness of the classification pipeline. This augmentation technique is applied to both the training sets for zoom group 3 and zoom group 1 (samples of different zoom levels are illustrated in Table 3.5). Furthermore, the limitations associated with the data augmentation based on SinGAN were also analysed.

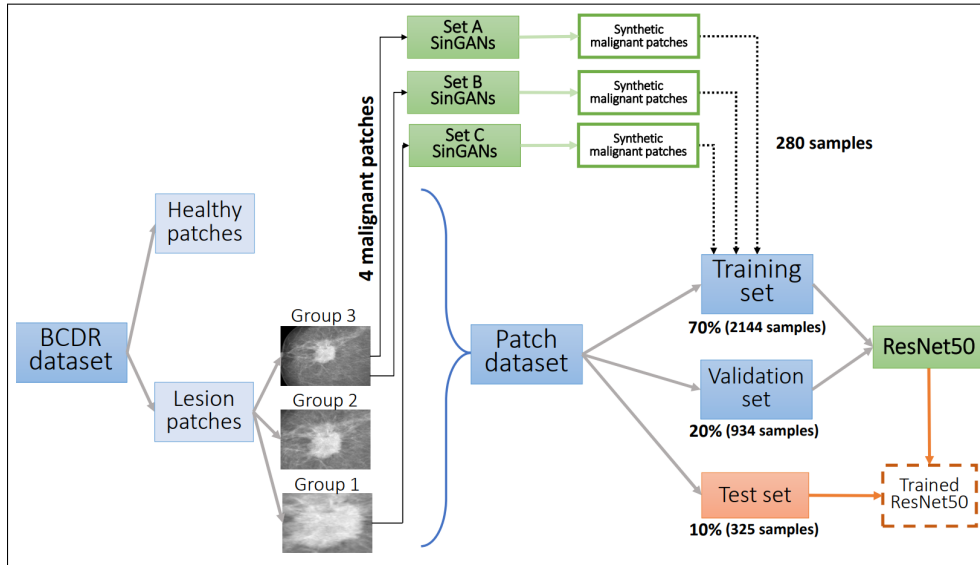


FIGURE 3.14: General pipeline of the experiments performed. Patches of healthy and lesion samples are initially extracted from the BCDR dataset (technique in Section 3.2.2). Three patches at three distinct levels of zoom are extracted for each lesion (Rationale in Section 3.2.1). The patch dataset is subsequently splitted into training, validation, and test sets, with three folds computed for all experiments. The number of patches depicted in the figure assumes the selection of a single group of zoom for the lesions. To augment the malignant class, different patches from Group 3 and Group 1 are chosen, and SinGAN models are trained on each of them. Synthetic patches are then generated from these models and incorporated into the training dataset. Sets A, B, and C, each referring to four SinGAN models, are named to represent this process, with further details provided in Figure 3.12.



## Chapter 4

# Results and discussion

### 4.1 Binary experiment

The first experiment performed was the binary task, where the target label indicated whether the patch represented healthy tissue or a suspicious lesion, with the positive class representing the healthy tissue. This experiment was conducted using the group of zoom 3 for the lesions, as the presence of more healthy adjacent tissue around the lesion could make it harder to detect.

The results for this experiment, conducted over 3 folds, demonstrated a test accuracy of  $0.924 \pm 0.009$  and a test AUC of  $0.971 \pm 0.009$  (Table 4.1). A high AUC metric suggests that the model is effective at discriminating between classes and ranking instances. However, the accuracy of individual classifications is slightly lower than the AUC, which may be attributed to potential misclassifications near the decision boundary. In addition, it is important to note that the standard deviation is low, what demonstrates consistency against different folds, reinforcing the model's stability and reliability. For a better visualisation of the model's performance, the Receiver Operating Characteristic (ROC) curve is presented in Figure 4.1. The ROC curve showcases the trade-off between the True Positive Rate and False Positive Rate.

To further address the problem, the subsequent sections of the paper focused on converting the task into a multiclass classification problem. This enabled the classification of suspicious patches into benign or malignant lesions, providing a more detailed analysis of the detected abnormalities.

<b>Experiment</b>	<b>Accuracy (Mean <math>\pm</math> Standard Deviation)</b>	<b>AUC (Mean <math>\pm</math> Standard Deviation)</b>
Binary	$0.924 \pm 0.009$	$0.971 \pm 0.009$

TABLE 4.1: Performance test metrics for Binary Experiment. The table presents the performance metrics of the ResNet model in the binary task of classifying suspicious lesions. The AUC metric is higher than the Accuracy metric, suggesting that the model is more effective at discriminating between classes and ranking instances rather than achieving perfect individual classification accuracy. The low standard deviation indicates consistent performance across different folds, reinforcing the model's stability and reliability.

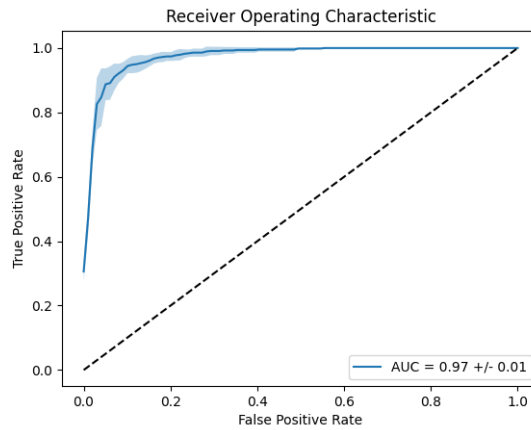


FIGURE 4.1: Receiver Operating Characteristic (ROC) Curve for Binary Experiment at testing. The shaded area surrounding the ROC curve represents the standard deviation across different folds, indicating the variability in performance. In this experiment, the ResNet model achieved a significant AUC value, reinforcing its capability to classify suspicious lesions effectively.

## 4.2 Training at different zoom levels of the lesions

In this section, the results of a set of experiments are presented. The patches from lesions were extracted, considering different percentages of healthy adjacent tissue and grouped into three levels of zoom, as shown in Figure 3.5. These grouping strategies served two purposes: (a) to explore the model’s robustness against different window sizes, aiming for implementation in a sliding windows procedure for lesion detection, and (b) to handle variations in the precision or consistency of the annotation masks provided by radiologists. These explorations aimed to evaluate the model’s dependency on precise annotations and its ability to generalise well in real-world scenarios where annotations may be less precise or consistent.

Each of the performed experiments involved training the classifier on one group of zoom levels and testing on samples from all the groups. It is important to note that each lesion was exclusively present in one of the train-validation-test sets. For instance, if the training set contained samples from group 1, those lesions were not present in the test set for any of the zoom level groups. Furthermore, an additional experiment was conducted by training the model using lesion patches from all three groups.

Table 4.2 presents the accuracy obtained for this set of experiments. The highest accuracy was achieved when training the model using lesion patches from all three groups of zoom levels. However, it is important to note that the accuracy obtained in the multiclass classification experiments is slightly lower than that of the binary problem. Conversely, the lowest accuracy was observed when training the model exclusively on group 1 of zoom levels, which corresponds to lesions with the least adjacent tissue. This finding suggests that the zoom level of the patch plays a significant role in the classifier’s performance when tested on lesion patches at different levels of zoom.

Additionally, Figure 4.2 illustrates the AUC metric for the different classes in the experiments. Across all experiments, the malignant class consistently exhibits

Experiment		Accuracy (Mean $\pm$ Standard Deviation)
Train-Val	Test	
1:2:3	1:2:3	$0.837 \pm 0.045$
1	1:2:3	$0.748 \pm 0.044$
2	1:2:3	$0.804 \pm 0.035$
3	1:2:3	$0.807 \pm 0.035$

TABLE 4.2: Accuracies of the experiments performed on Figure 4.2 training the classifier only on one group of zoom levels and testing on samples from all the groups. The majority of experiments show comparable accuracy values, except for the second experiment, which also exhibits the lowest performance in terms of the AUC metric in Figure 4.2.

the lowest AUC value, indicating that the classifier performs comparatively worse for this class. These results highlight the challenge in accurately classifying malignant lesions and suggest the need for further investigation and improvement in the model’s performance for this specific class. The corresponding ROC curves for each of the experiments, are included in the Appendix 6.2.1, in Figure 6.3.

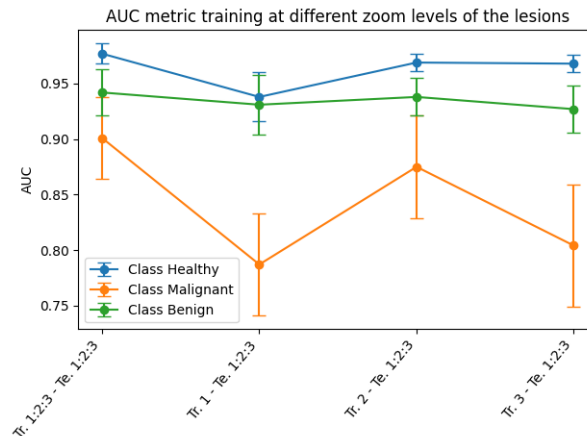


FIGURE 4.2: Area under ROC curve of the experiments performed training only on one group of zoom levels and testing on samples from all the groups. The malignant class consistently exhibits the lowest AUC value.

It is worth highlighting the results of the third experiment, where the model was trained on group 2 of zoom levels. This experiment demonstrates superior performance compared to training exclusively on group 1 or group 3. A hypothesis can be formulated that training on the group of zoom 2 represents an intermediate point that optimally captures the necessary features for later testing on groups 1 and 3. Conversely, training exclusively on set 1 or set 3 and subsequently testing on all three sets leads to poorer performance. This observation further supports the notion that the tightness of the bounding box around the lesion plays a significant role in the classifier’s performance.

### 4.3 Analysing classifier robustness against variations in the quality or accuracy of the annotation mask

With the purpose of analysing how variations in the quality of the annotation mask of a lesion influences classifier performance, this section presents experiments conducted by training and testing on different combinations of zoom level groups. Table 4.3 provides an overview of the accuracy obtained for this set of experiments at testing time. The highest accuracies were achieved when the model was tested on lesion patches from the same group of zoom levels the model was trained on. This consistency is also reflected in the AUC metric, as shown in Figure 4.3. However, the most interesting results were observed when the model was tested on a different level of zoom the model was trained on. The worst performance in terms of accuracy was observed when the classifier trained on group 1 was tested on group 3. In terms of the AUC, it is the experiment with the second worst result. Observing the AUC test results for the different classes, the worst result was for the classifier trained on group 3 tested on group 1. As is observed also in Figure 4.3 that when the model is trained on the intermediate level of zoom of group 2, the performance when testing on group 1 or 3 is better for the malignant class than training on group 1 or 3. The ROC curves of the different experiments are included in Figure 6.3 in Appendix 6.2.2.

Experiment		Accuracy (Mean $\pm$ Standard Deviation)	Malignant class AUC (Mean $\pm$ Standard Deviation)
Train-Val	Test		
1	1	0.929 $\pm$ 0.034	0.948 $\pm$ 0.043
2	2	0.899 $\pm$ 0.024	0.930 $\pm$ 0.020
3	3	0.887 $\pm$ 0.031	0.928 $\pm$ 0.036
3	1	0.865 $\pm$ 0.088	0.677 $\pm$ 0.076
2	1	0.877 $\pm$ 0.009	0.840 $\pm$ 0.041
1	3	0.780 $\pm$ 0.021	0.709 $\pm$ 0.066
2	3	0.835 $\pm$ 0.025	0.893 $\pm$ 0.052

TABLE 4.3: Accuracies at testing time of the experiments performed on Figure 4.3 and the AUC metric for the Malignant class. Training and testing the model on different combinations of zoom level groups.

### 4.4 SinGAN-based data augmentation

After the previous sets of experiments, it can be consistently concluded that the classifiers perform comparatively worse for the malignant class, which based on Figure 3.7a, it is the minority class. In this section, data augmentation with SinGAN models is explored aiming to improve the classifier performance for this specific class. The subsequent experiments involving SinGAN data augmentation focus on the experiment with the lowest AUC (Area Under the Curve) for the malignant class, as described in Section 4.3, which serves as the baseline.

This particular experiment aims to assess the performance of a classifier trained on the level of zoom of group 3, which utilises less accurate annotation masks of lesions when tested on group 1. In group 1, the annotation masks consist of tightly

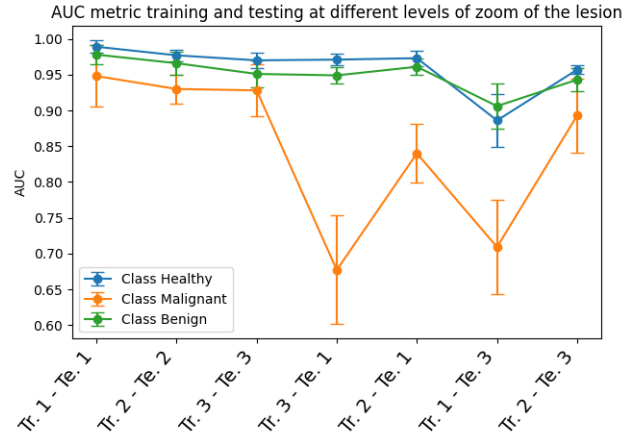


FIGURE 4.3: Area under the Receiver Operating Characteristic (ROC) Curve for the multiclass task training and testing the model on different combinations of zoom level groups.

bounding boxes around the lesions. The objective of this experiment is to determine whether training a classifier with a broader annotation mask of the lesion, which accounts for inter- and intra-observer variability, can accurately classify patches of lesions with fewer adjacent healthy tissue.

This analysis seeks to determine if the classifier has successfully learned the specific patterns of the lesion itself, thereby preventing scenarios similar to the well-known anecdote in the field of AI involving a neural network model trained to differentiate between wolves and huskies (Ribeiro, Singh and Guestrin, 2016). In that case, the model accurately classified test samples but had seemingly learned to predict "Wolf" if there was snow and "Husky" otherwise, disregarding animal features or patterns. Drawing upon this analogy, our analysis examines the performance of a classifier trained on the animal (or breast lesion) with its corresponding background (or healthy adjacent tissue), when subsequently tested on images that solely depict the animal body (or a tight bounding box around the lesion).

#### 4.4.1 Generation and evaluation images with SinGAN models

As introduced in Section 3.4.3, the Single Image FID (SiFID) metric (proposed by Shaham, Dekel and Michaeli, 2019) was used. The SiFID metric allows to quantitatively evaluate the similarity between the original and synthetic images, providing valuable insights into the performance of the SinGAN models in generating realistic and faithful images that capture the characteristics of the original data.

To establish a reference point for comparison, the SiFID between the original image and randomly generated noise images was also computed. This baseline value provided an indication of the level of dissimilarity expected between unrelated images. Furthermore, when comparing the original image with itself, a SiFID value of absolute 0 was obtained, indicating a perfect match between identical images.

The SiFID values for the generated synthetic images, as shown in Figure 4.4, were found to be very close to 0 with a small standard deviation. This implies that the feature distributions of the synthetic images closely resemble those of the original images, indicating a high degree of quality and fidelity in the image generation

process. The small standard deviation further suggests consistency across different images of the same set.

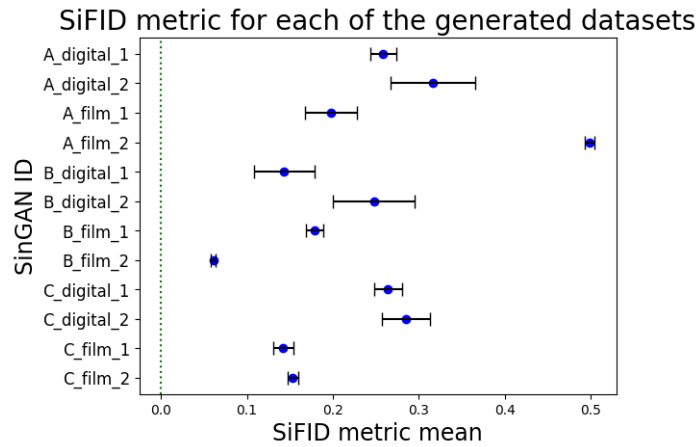


FIGURE 4.4: Distribution of SiFID metric values for datasets generated using a SinGAN model trained on different single images. The y-axis represents the ID of the training image used to generate each dataset, indicating the set (A, B, or C) and the format (digital or film). The dots represent the mean value, and the horizontal black line represents the standard deviation. For reference, the metric was computed with respect to a dataset of random noise images as an upper-bound, which had a SiFID of  $39.282 \pm 0.143$ , so the metrics obtained demonstrate the excellent quality of the generated images."

#### 4.4.2 Improving baseline performance with SinGAN-based data augmentation

In order to investigate the impact of SinGAN data augmentation on the malignant class within the training data, additional samples were incorporated into the dataset. These samples were generated using four SinGAN models, two from digital patches and two from film patches. This approach allowed the exploration of the effects of data augmentation across different formats. Subsequently, the classifier trained with this augmented dataset was tested on group 1 to evaluate its performance compared to the baseline.

The data augmentation was performed on generated samples from four patches of group 3, corresponding to Set B in Figure 3.12, as well as on augmented samples from patches of group 1, corresponding to Set C in the same Figure. For consistency, the same lesions were used in both Sets but at different levels of zoom. These two experiments were labeled as i and ii, and the accuracy metric is presented in Table 4.4, while the AUC for the different classes can be observed in Figure 4.5.

The performance of the classifier for the malignant class exhibits notable improvements in terms of the AUC, increasing from  $0.677 \pm 0.076$  to  $0.718 \pm 0.044$ , when augmenting the training data with SinGAN-generated samples from the group of zoom 3. The accuracy also reflect an enhancement, which remains consistent at  $0.865 \pm 0.008$ . In experiment (ii) (Table 4.4), where the training set is augmented with samples from SinGAN models trained on zoom 1 samples, the performance boost is even more substantial. The AUC for the malignant class rises to  $0.771 \pm 0.045$  and

the accuracy, to  $0.881 \pm 0.007$ . However, it is important to note that the test performance is not as favorable when directly testing on the group of zoom 3. This positive outcome highlights the potential of data augmentation with synthetically generated images based on SinGAN models, presenting a promising result. The ROC curves for data augmentation and non data augmentation experiments are included in Figure 4.6.

Experiment	Set of models	Training image IDs	Zoom group
i	B	13, 77, 2789, 3941	3
ii	C	11, 75, 2787, 3939	1

TABLE 4.4: Specification of the SinGAN models (Table 3.12) used for the experiments performed in Figure 4.5.

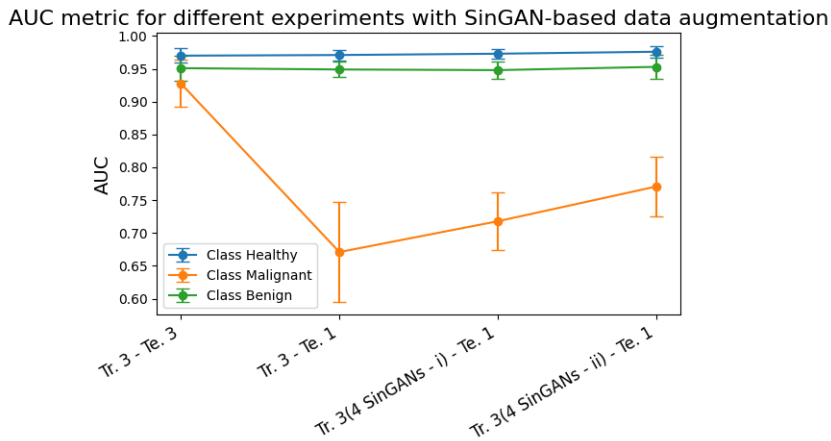


FIGURE 4.5: AUC metric for each class for different experiments. In the first pair of experiments, data augmentation is not used and only the test (Te.) set is changed. In the second pair of experiments, SinGAN-based data augmentation was used for two sets of 4 SinGAN models, specified in Table 4.4.

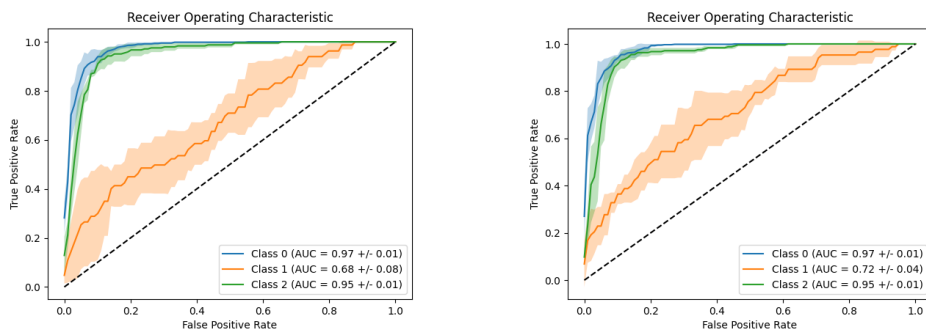
Experiment		Accuracy (Mean $\pm$ Std. Deviation)	Malignant class AUC (Mean $\pm$ Std. Deviation)
Train-Val	Test		
3	3	$0.887 \pm 0.031$	$0.928 \pm 0.036$
3	1	$0.865 \pm 0.008$	$0.677 \pm 0.076$
3(4 SinGANs - i)	1	$0.870 \pm 0.016$	$0.718 \pm 0.044$
3(4 SinGANs - ii)	1	$0.881 \pm 0.007$	$0.771 \pm 0.045$

TABLE 4.5: Accuracies and AUC metrics for the malignant class of the experiments performed on Figure 4.5.

Additionally, misclassification rates for the test set were computed. This involved calculating the ratio of misclassified samples with a specific feature to the total number of samples with that feature in the test set. In relation to the calcification feature presented in Figure 3.8, a surprising outcome was observed in these experiments. When testing for group 1 without data augmentation, the misclassification rate of samples with calcification was found to be  $0.248 \pm 0.104$ . However,

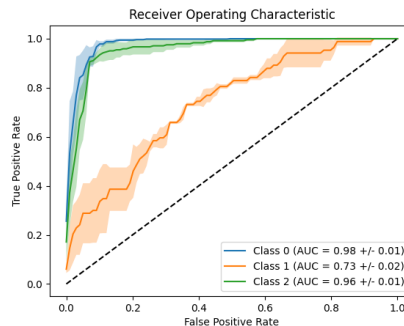


when augmenting the dataset with four SinGAN models in experiment (i), this result decreased to  $0.238 \pm 0.064$ . It is worth noting that two of the SinGAN models used in this experiment were trained on calcification samples (ID 2789 and 3941 in Table 3.12). As discussed in Section 3.3, according to Nalawade, 2009, the appearance of absolutely benign calcifications on mammography may render biopsy unnecessary. These findings highlight promising slight improvements in the consistency of the classifier using SinGAN-based data augmentation, particularly in identifying calcifications. Furthermore, for experiment (ii), where the Set C (Figure 3.12) of SinGAN models was used, corresponding to the same lesions as in the Set B used in experiment (i), the misclassification rate for samples with calcification was further improved to  $0.211 \pm 0.074$ . This outcome is consistent with the initial hypothesis that augmenting the data with synthetic images generated from patches of calcifications could decrease the misclassification rate for calcifications.



(A) Training only on group 3 and testing on group 1.

(B) Augmenting training samples with synthetic samples of group of zoom 3.



(C) Augmenting training samples with synthetic samples of group of zoom 1.

FIGURE 4.6: Receiver Operating Characteristic (ROC) Curve for the multiclass task training on level of zoom group 3, augmenting the training data with synthetic samples of group 3 (Set B in 3.12) and of group 1 (Set C in the same Figure). The shaded area corresponds to the standard deviation for the different k-folds for which the experiment was done.

#### 4.4.3 Synthetic dataset generated from a different number of SinGAN models

After the promising results obtained using single image generative models for data augmentation, the interest in exploring SinGAN opportunities led to the set of experiments explored in this section. The aim was to investigate the potential benefits of



incorporating diversity into the training set through synthetic samples. The number of synthetic samples remained constant across all experiments, while the number of SinGAN models used to generate those samples varied. Different synthetic datasets, which augmented the training set by the same size, were created by combining different subsets of the 12 SinGAN models. Each SinGAN model was trained on one of the 12 images presented in Figure 3.12. This process was specifically carried out for SinGAN models trained on samples from the group of zoom 3. The IDs of the training images used for generating each synthetic dataset are provided in Figure 4.6. The accuracy achieved in each experiment is presented in Table 4.7, while the AUC values for the three classes are plotted in Figure 4.7. All the ROC curves for each experiment and for each class are included in the Appendix 6.2.3, in Figure 6.5.

The results of the entire set of experiments indicate that the performance without data augmentation is worse in terms of both AUC and accuracy. This observation confirms that data augmentation with generated samples using SinGAN models consistently improves the performance and robustness of the model across different levels of zoom.

However, there is an important observation to be made from these results. While the experiments utilising synthetic samples from 2 and 4 SinGAN models demonstrate a considerable and promising enhancement, the performance is substantially poorer when using 6 or 8 SinGAN models to generate the synthetic dataset. This observation suggests that the choice of the training image for SinGAN sample generation has a notable influence on the model’s performance.

As a final observation regarding the misclassification rate for calcifications mentioned in Section 4.4.2, it was observed that augmenting the dataset with images generated from samples of lesions with calcification resulted in an improvement in this rate. However, in experiment (iv) of this section, a contrasting effect was observed. None of the SinGAN models from Set A (3.12) were trained on a calcification lesion patch, and these models were used to generate the synthetic dataset for this experiment. Surprisingly, the misclassification rate for calcifications did not improve and, in fact, worsened from  $0.248 \pm 0.103$  to  $0.261 \pm 0.112$ .

Experiment	Set of models	Training image IDs	Zoom group
i	B	77, 2789	3
ii	B	13, 3941	3
iii	B	13, 77, 2789, 3941	3
iv	A	51, 168, 2180, 3884	3
v	A&B	13, 77, 168, 2789, 3941, 2180	3
vi	A&B	13, 77, 51, 2789, 3941, 3884	3
vii	A&B	77, 51, 168, 2789, 3884, 2180	3
viii	A&B	13, 51, 168, 3941, 3884, 2180	3
ix	A&B	13, 51, 168, 77, 3941, 3884, 2180, 2789	3

TABLE 4.6: Specification of the image IDs (Table 3.12) used for training the SinGAN models used for the experiments performed in Figure 4.7.

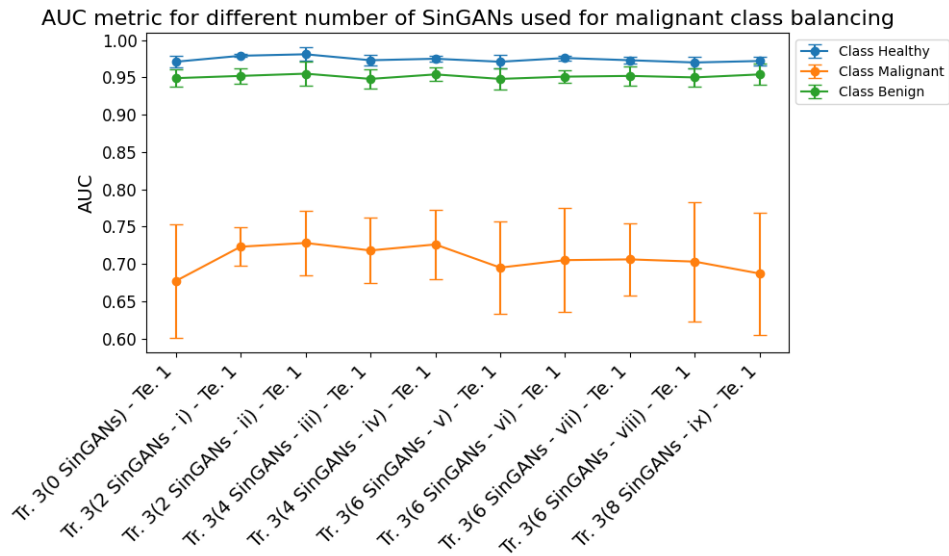


FIGURE 4.7: AUC metric for each class for different experiments. For all the experiments the classifier was trained (Tr.) on group 3 of zoom and tested (Te.) on group 1. Each SinGAN model set is specified in Table 4.6.

Experiment		Accuracy	Malignant class AUC
Train-Val	Test	(Mean $\pm$ Std. Deviation)	(Mean $\pm$ Std. Deviation)
3(0 SinGANs)	1	0.865 $\pm$ 0.008	0.677 $\pm$ 0.076
3(2 SinGANs - i)	1	0.891 $\pm$ 0.011	0.723 $\pm$ 0.026
3(2 SinGANs - ii)	1	0.892 $\pm$ 0.004	0.728 $\pm$ 0.043
3(4 SinGANs - iii)	1	0.870 $\pm$ 0.016	0.718 $\pm$ 0.044
3(4 SinGANs - iv)	1	0.873 $\pm$ 0.012	0.726 $\pm$ 0.046
3(6 SinGANs - v)	1	0.874 $\pm$ 0.005	0.695 $\pm$ 0.062
3(6 SinGANs - vi)	1	0.880 $\pm$ 0.009	0.705 $\pm$ 0.070
3(6 SinGANs - vii)	1	0.879 $\pm$ 0.014	0.706 $\pm$ 0.048
3(6 SinGANs - viii)	1	0.869 $\pm$ 0.018	0.703 $\pm$ 0.080
3(8 SinGANs - ix)	1	0.882 $\pm$ 0.012	0.687 $\pm$ 0.082

TABLE 4.7: Accuracies of the experiments performed on Figure 4.7, the index of which SinGAN models are used for generating the synthetic set in each experiment is included in Table 4.6.

## Chapter 5

# Conclusions and further work

In this project, a Deep Learning model has been implemented to classify patches extracted from mammograms into three classes: healthy (no lesion), benign, and malignant. These are biopsy results of the lesions and provide a reliable ground truth for training the model. This methodology allowed the assessment of multiple regions within an image and predictions based on individual patches. Furthermore, it was observed that the performance for the malignant class was relatively poorer. To enhance the robustness and generalisability of the classification model, patches with lesions were extracted at various levels of zoom, representing different quantities of adjacent healthy tissue. This technique aimed to simulate inter- and intra-observer variability in expert annotated masks or sliding window detection software with different window sizes. The influence of the adjacent tissue of the lesion to the performance of the classifier has been consistently proved.

For the subsequent experiments, the performance of a classifier trained on patches with less zoom was analysed when tested on patches with more zoom and minimal adjacent healthy tissue. This approach aimed to assess the classifier's ability to learn the patterns of the lesions themselves. A SinGAN-based data augmentation methodology was applied to evaluate its potential for balancing the malignant class, and promising results were obtained. Additionally, the impact of diversity in the synthetic dataset on performance was analysed, revealing an unexpected trend of performance improvement. However, the improvement relative to the baseline was consistently observed. The hypotheses explored throughout the project and their corresponding conclusions are summarised in Table 5.1.

As stated in the introduction section, to the best of our knowledge, this project represents a novel study in the application of single image generative models for breast cancer diagnosis. Given its novelty, it opens up several avenues for further analysis, validation, and implementation of derived tools. Suggestions for future work and motivations to continue this study can be categorised into two sections: (a) improvements of the techniques employed and (b) potential clinical applications.

Regarding the improvement of the employed techniques, it is recommended to augment the dataset from different sources, considering that this project only utilised the BCDR dataset. Restricting the dataset limits the representation of the population and, for practical implementation in medical applications, it is crucial to avoid distribution shifts. In terms of patch extraction, randomising the zoom of healthy patches, similar to what was done with the lesions, can provide a more comprehensive analysis. However, the focus was primarily on the lesions, as they were the critical point for the classification task. Additionally, incorporating a more complex ResNet or exploring alternative classification models could enhance the analysis.

#	Hypothesis	Result	✓/✗
1	The malignant class exhibits inferior performance compared to other classes.	Consistently supported across all conducted experiments. Moreover, it was observed that the malignant class represents the minority within the dataset.	✓
2	A classifier trained exclusively on lesions with a broad bounding box (group 3 of zoom) is expected to exhibit lower performance when tested on lesions with minimal adjacent tissue (group 1 of zoom) compared to when tested on lesions from group 3.	Experiments in section 4.3 confirm the hypothesis. It seems more difficult for the classifier to learn the specific patterns of the lesion itself when trained solely on lesions with more surrounding tissue (group 3).	✓
3	The performance of a classifier trained on lesions from group 3 (broader bounding box) improves when tested on lesions from group 1 (tight bounding box) if the training data is augmented with synthetic malignant lesions from group 3 of zoom.	The hypothesis was confirmed by experiment (i) conducted in section 4.4.2.	✓
4	Augmenting the data with synthetic images generated from patches of calcifications will decrease the misclassification rate for calcifications. However, if the synthetic dataset is not generated from calcification lesions, it will not lead to any improvement.	It was confirmed by experiments (i) and (ii) in Section 4.4.2, where synthetic samples generated from calcification patches improved the baseline. However, in experiment (iv) in Section 4.4.3, where no calcification patches were used for SinGAN data augmentation, the misclassification rate did not improve.	✓
5	The improvement of the performance stated in Hypothesis 3 is larger when augmenting the training data with synthetic lesions from group 3.	The hypothesis was confirmed by experiment (ii) conducted in Section 4.4.2.	✓
6	The larger the number of SinGAN models used for generating the synthetic dataset, the greater the enhancement of performance for the malignant class.	Although the experiments carried out in Section 4.4.3 showed a consistent enhancement in the classification performance for the malignant class, the expected improvement when augmenting data with more SinGANs was often minimal. Therefore, having more SinGANs is not always better.	✗

TABLE 5.1: Overview of the hypotheses explored in the project development with the corresponding outcomes and conclusions.

Moreover, conducting more folds for each experiment would allow for performing robust significance statistical tests. Due to the significant training time required for each classifier (approximately 3.5 hours), only three folds were computed, which limited the statistical analysis.

Concerning SinGAN augmentation, conducting further analysis on the rejected hypotheses in Table 5.1 would be beneficial. It would be valuable to investigate the patterns that contributed to a poorer enhancement for the of certain SinGAN models used in experiments of section 4.4.3. It can be also explored other single-image generative models, such as SinFusion model proposed by Nikankin, Haim and Irani, 2022. Additionally, revisiting the SinGAN model itself could be considered. In their Con-SinGAN paper, Hinz et al., 2021 proposed promising enhancements to the original SinGAN technique. Unlike the original structure, which fixed each GAN once trained, Hinz et al., 2021 suggested training multiple stages concurrently in a sequential multi-stage manner. This modification enables better capture of semantic structure between patches, resulting in models with fewer stages and faster training (up to 6 times faster).

On the other hand, considering potential clinical applications, the involvement of medical experts is essential. The use of a Deep Learning model, as demonstrated in this study, offers a pre-biopsy result that can provide valuable insights to healthcare professionals, helping them in making informed decisions regarding subsequent diagnostic procedures and treatment plans. Two specific applications were considered during the project's development. Firstly, the classification task could be integrated into a lesion detection software using a sliding windows procedure. Secondly, an API could be developed, enabling radiologists to submit selected suspicious areas (bounding boxes) from mammograms and promptly receive a measure indicating the likelihood of a benign or malignant lesion.

For both application, the classifier should be robust against variations in the lesion zoom levels and the accuracy of the annotation masks. The model should be adapted to different window sizes in the detection software and varying accuracies of bounding boxes submitted to the Application Programming Interface (API), considering inter-observer variability among different radiologists or even intra-observer variability between different examinations conducted by the same radiologist.

These two applications have the purpose of helping to assist on decision making, in general, and for a biopsy intervention decision and its urgency, in particular. As highlighted by Shyamala, Girish and Murgod, 2014, biopsy interventions carry inherent risks for patients, such as the possibility of dislodging and seeding of neoplastic altered cells. It is important to emphasise that any clinical implementation requires a comprehensive understanding of the healthcare system and diagnostic protocols. It is evident that the involvement of clinical experts is essential for future work on this project. The first step would be to gain a deeper understanding of the actual challenges, which would lead to the development of effective solutions. Additionally, it is important to validate the realism of the generated synthetic images through expert assessment. Furthermore, the previously presented applications could potentially be developed for utilisation by radiologists. Therefore, the validation of these tools should be conducted by healthcare experts. In conclusion, the participation of clinical experts in the study is crucial to ensure the wider adoption of these AI-tools within the healthcare system.



# Bibliography

- Abdelrahman, Leila et al. (2021). 'Convolutional neural networks for breast cancer detection in mammography: A survey'. In: *Computers in biology and medicine* 131, p. 104248.
- Arevalo, John et al. (2015). 'Convolutional neural networks for mammography mass lesion classification'. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 797–800. DOI: [10.1109/EMBC.2015.7318482](https://doi.org/10.1109/EMBC.2015.7318482).
- Bi, Wenya Linda et al. (2019). 'Artificial intelligence in cancer imaging: clinical challenges and applications'. In: *CA: a cancer journal for clinicians* 69.2, pp. 127–157.
- Breast Imaging Reporting & Data System (BI-RADS®) (2023). URL: <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads>.
- Canadian Cancer Society (2023). URL: <https://cancer.ca/en/treatments/tests-and-procedures/mammography/breast-density> (visited on 01/06/2023).
- Elmore, Joann G. et al. (1994). 'Variability in Radiologists' Interpretations of Mammograms'. In: *New England Journal of Medicine* 331.22. PMID: 7969300, pp. 1493–1499. DOI: [10.1056/NEJM199412013312206](https://doi.org/10.1056/NEJM199412013312206). eprint: <https://doi.org/10.1056/NEJM199412013312206>. URL: <https://doi.org/10.1056/NEJM199412013312206>.
- Eric et al. (2018). 'Conditional infilling GANs for data augmentation in mammogram classification'. In: *Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 3*. Springer, pp. 98–106.
- Goodfellow, Ian et al. (2020). 'Generative adversarial networks'. In: *Communications of the ACM* 63.11, pp. 139–144.
- Guevara Lopez, Miguel Angel et al. (Jan. 2012). 'BCDR: A BREAST CANCER DIGITAL REPOSITORY'. In: pp. 1065–1066.
- Gulrajani, Ishaan et al. (2017). 'Improved training of wasserstein gans'. In: *Advances in neural information processing systems* 30.
- He, Kaiming et al. (2016). 'Identity mappings in deep residual networks'. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pp. 630–645.
- Heusel, Martin et al. (2017). 'GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf).
- Hinz, Tobias et al. (2021). 'Improved techniques for training single-image gans'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1300–1309.
- Hu, Xiaodan et al. (2018). 'Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks'. In: *arXiv preprint arXiv:1811.05817*.



- LeCun, Yann, Yoshua Bengio et al. (1995). 'Convolutional networks for images, speech, and time series'. In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.
- Lee, Juhun and Robert M. Nishikawa (2022). 'Identifying Women With Mammographically-Occluded Breast Cancer Leveraging GAN-Simulated Mammograms'. In: *IEEE Transactions on Medical Imaging* 41.1, pp. 225–236. DOI: [10.1109/TMI.2021.3108949](https://doi.org/10.1109/TMI.2021.3108949).
- Lekadir, Karim et al. (2021). 'Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging'. In: *arXiv preprint arXiv:2109.09658*.
- Nalawade, Yojana V (2009). 'Evaluation of breast calcifications.' In: *The Indian journal of radiology imaging* 19.4, 282–286. DOI: <https://doi.org/10.4103/0971-3026.57208>.
- Nikankin, Yaniv, Niv Haim and Michal Irani (2022). 'SinFusion: Training Diffusion Models on a Single Image or Video'. In: *arXiv preprint arXiv:2211.11743*.
- Official pytorch implementation of the paper: "SinGAN: Learning a Generative Model from a Single Natural Image" (2019). (Visited on 30/05/2023).
- OpenCV: Open Source Computer Vision Library (n.d.). URL: <https://opencv.org/> (visited on 01/06/2023).
- Osuala, Richard et al. (2023a). 'Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging'. In: *Medical Image Analysis* 84, p. 102704. ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2022.102704>. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522003322>.
- Osuala, Richard et al. (2023b). 'medigan: a Python library of pretrained generative models for medical image synthesis'. In: *Journal of Medical Imaging* 10.06. DOI: [10.1117/1.jmi.10.6.061403](https://doi.org/10.1117/1.jmi.10.6.061403). URL: <https://doi.org/10.1117/2F1.jmi.10.6.061403>.
- PyTorch: models and pre-trained weights (2023). URL: <https://pytorch.org/vision/stable/models.html> (visited on 01/06/2023).
- PyTorch: Weighted Random Sampler (2023). URL: <https://pytorch.org/docs/stable/data.html#torch.utils.data.WeightedRandomSampler> (visited on 01/06/2023).
- Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin (2016). "'Why should i trust you?' Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
- Shaham, Tamar Rott, Tali Dekel and Tomer Michaeli (2019). 'SinGAN: Learning a Generative Model From a Single Natural Image'. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4569–4579. DOI: [10.1109/ICCV.2019.00467](https://doi.org/10.1109/ICCV.2019.00467).
- Shyamala, K, HC Girish and Sanjay Murgod (2014). 'Risk of tumor cell seeding through biopsy and aspiration cytology'. In: *Journal of International Society of Preventive & Community Dentistry* 4.1, p. 5.
- Szafranowska, Zuzanna et al. (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. SPIE, pp. 169–177.
- Thambawita, Vajira et al. (2022). 'SinGAN-Seg: Synthetic training data generation for medical image segmentation'. In: *PloS one* 17.5, e0267976.
- WHO: Breast Cancer Fact Sheet (2021). URL: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (visited on 30/05/2023).
- WHO: Cancer Fact Sheet, (2022). URL: <https://www.who.int/news-room/fact-sheets/detail/cancer> (visited on 30/05/2023).
- Wu, Eric, Kevin Wu and William Lotter (2020). *Synthesizing lesions using contextual GANs improves breast cancer classification on mammograms*. arXiv: 2006.00086 [eess.IV].



## Chapter 6

# Appendix

### 6.1 ResNet model graph

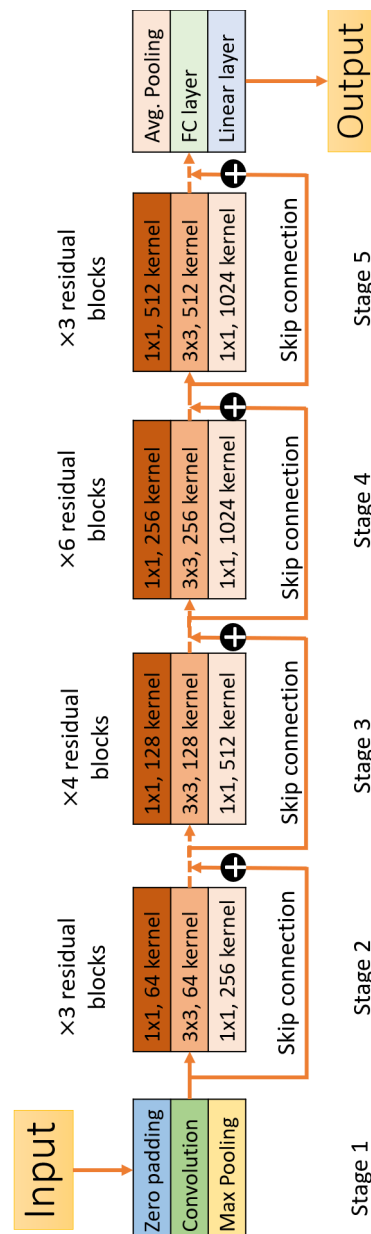
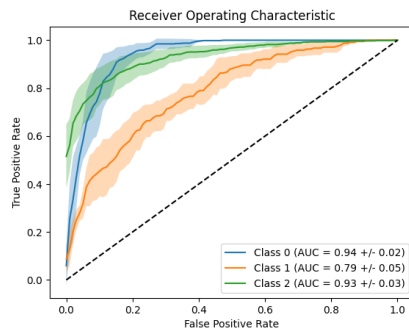


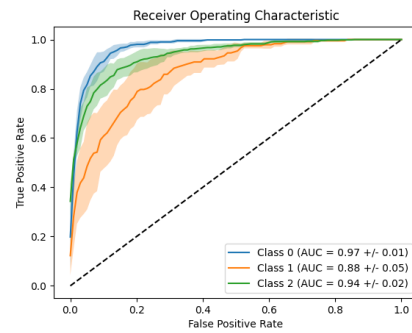
FIGURE 6.1: Schema of the model used for the classification task in this project. More details explained in Section 3.5.

## 6.2 ROC curves

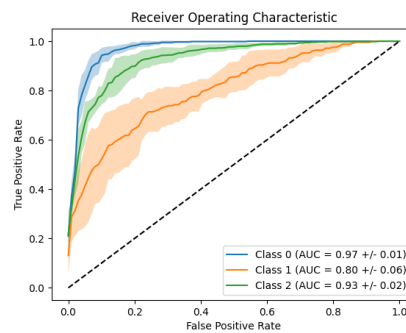
### 6.2.1 ROC curves of experiments computed in section 4.2



(A) Training only on group 1 and testing on all of them.



(B) Training only on group 2 and testing on all of them.



(C) Training only on group 3 and testing on all of them.

FIGURE 6.2: Receiver Operating Characteristic (ROC) Curve for the multiclass task training on three different levels of zoom of lesions (Figure 3.5) and testing with all of them.

## 6.2.2 ROC curves of experiments computed in section 4.3

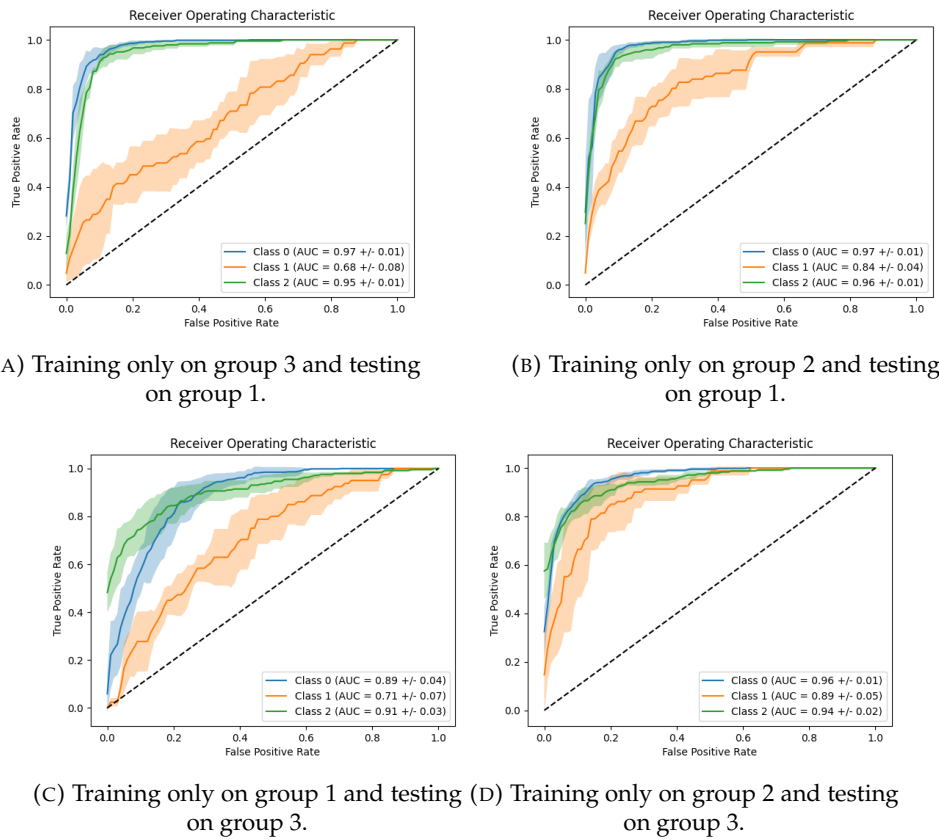
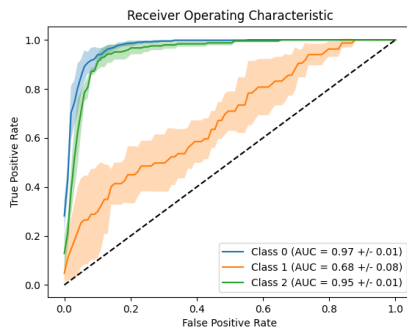
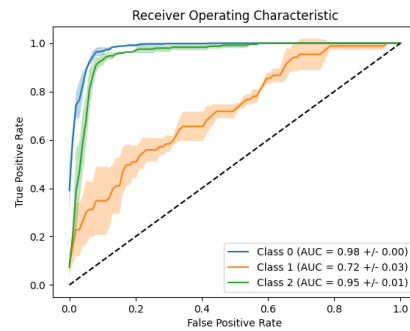


FIGURE 6.3: Receiver Operating Characteristic (ROC) Curve for the multiclass task training on the three different levels of zoom of lesions separately (Figure 3.5) and testing on a different level of zoom. The shaded area corresponds to the standard deviation for the different k-folds for which the experiment was done.

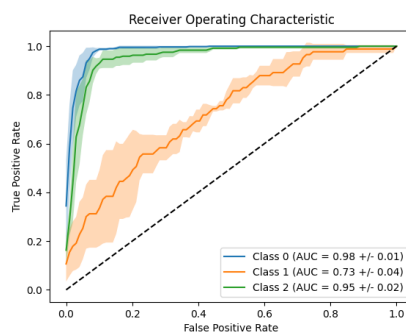
### 6.2.3 ROC curves of experiments computed in section 4.4.3



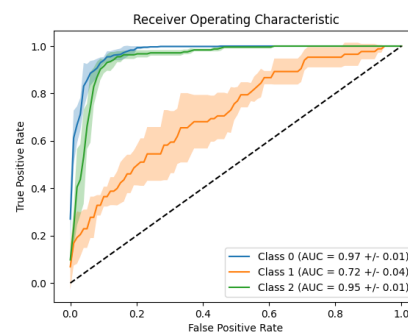
(A) Training the classifier on group 3 and testing on 1



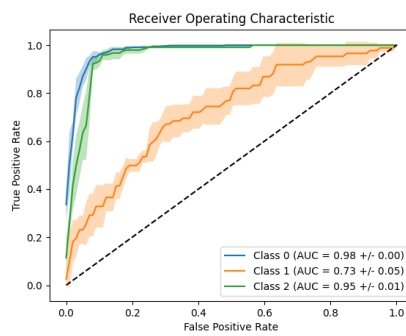
(B) Experiment (i) of Section 4.4.3 with 2 SinGANs.



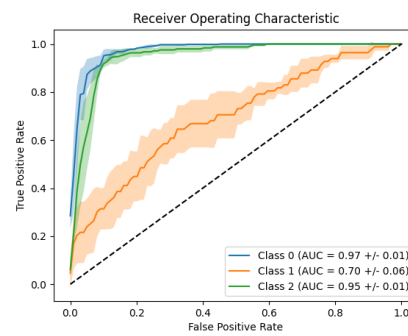
(C) Experiment (ii) of Section 4.4.3 with 2 SinGANs.



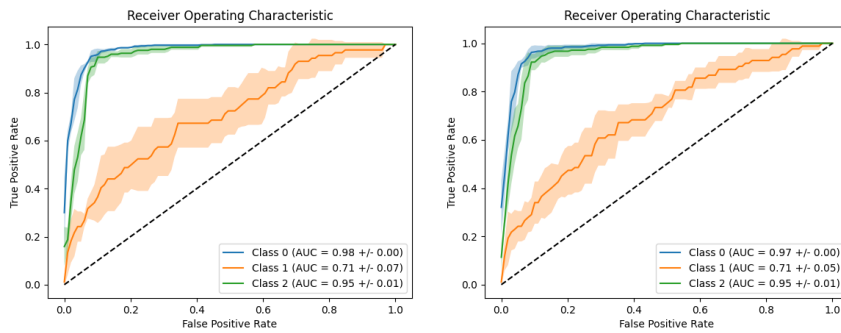
(D) Experiment (iii) of Section 4.4.3 with 4 SinGANs



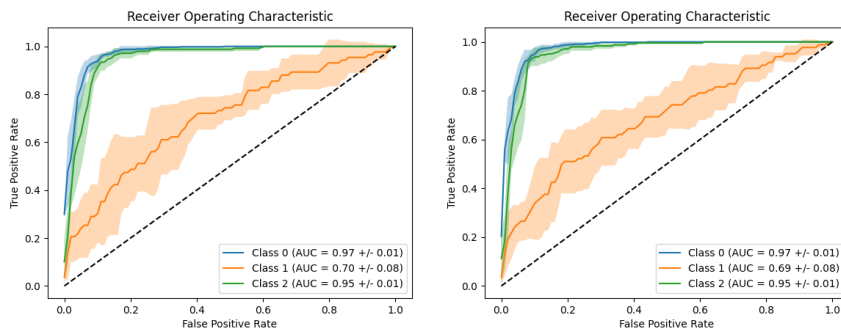
(E) Experiment (iv) of Section 4.4.3 with 4 SinGANs.



(F) Experiment (v) of Section 4.4.3 with 6 SinGANs.



(A) Experiment (vi) of Section 4.4.3 with 6 SinGANs. (B) Experiment (vii) of Section 4.4.3 with 6 SinGANs.



(C) Experiment (viii) of Section 4.4.3 with 6 SinGANs. (D) Experiment (ix) of Section 4.4.3 with 8 SinGANs.

FIGURE 6.5: Receiver Operating Characteristic (ROC) Curve for the multiclass task training on level of zoom group 3 and testing on group 1, presented in section 4.4.3. The training set is augmented with synthetic samples of group 3 (3.12) from a different number of SinGAN models. The shaded area corresponds to the standard deviation for the different k-folds for which the experiment was done.