UNIVERSITAT DE BARCELONA

Efficient Deep Learning for Medical Imaging: Precision Segmentation and Beyond

Lucas Martín Gago

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento 4.0. España de Creative Commons***.

This doctoral thesis is licensed under the ***Creative Commons Attribution 4.0. Spain License***.

EFFICIENT DEEP LEARNING FOR MEDICAL IMAGING: PRECISION SEGMENTATION AND BEYOND

LUCAS MARTÍN GAGO

UNIVERSITAT DE BARCELONA

UNIVERSITAT DE BARCELONA

Doctoral Program in Mathematics and Computer Science

EFFICIENT DEEP LEARNING FOR MEDICAL IMAGING: PRECISION SEGMENTATION AND BEYOND

Lucas Martín Gago

Supervisors

Dr. Laura Igual

Dr. Beatriz Remeseiro

Tutorship

Dr. Laura Igual

September 2025

UNIVERSITAT DE BARCELONA

Doctorat en Matemàtiques i Informàtica

APRENENTATGE PROFUND EFICIENT PER A LA IMATGE MÈDICA: SEGMENTACIÓ DE PRECISIÓ I MÉS ENLLÀ

Lucas Martín Gago

Supervisió

Dra. Laura Igual

Dra. Beatriz Remeseiro

Tutora

Dra. Laura Igual

Setembre 2025

Abstract

This thesis advances automated medical image analysis by introducing four deep learning frameworks that systematically address core technical barriers to clinical deployment, including computational efficiency, variability in image quality, and the robust integration of imaging with clinical data. Through a compendium of research articles, we develop state-of-the-art solutions spanning ultrasound and MRI.

First, we present an end-to-end framework for carotid intima-media thickness (CIMT) measurement in ultrasound images, achieving state-of-the-art atherosclerotic plaque characterization while delivering a 20× speed improvement (0.79 to 0.04 seconds per image). The system provides comprehensive outputs, including segmentation masks, automated measurements, and binary plaque detection, eliminating domain-specific post-processing requirements.

Secondly, leveraging the features extracted by our end-to-end model, we pioneer their integration into clinical survival models, demonstrating that learned imaging biomarkers significantly enhance cardiovascular risk stratification with a 20% improvement in patient risk reclassification beyond traditional clinical variables.

Third, we develop a multilevel EfficientNet-UNet++ architecture for 3D carotid vessel wall segmentation in black-blood MRI that achieves state-of-the-art performance through contextual slice concatenation and resolution optimization. The framework demonstrates optimal performance at 256×256 input resolution (6× original size) while maintaining computational efficiency through targeted multilevel processing.

Finally, we introduce a quality-aware segmentation framework with custom loss functions for explicit quality modeling during training. When applied to ultrasound colon wall segmentation, this approach achieves a 20% improvement on medium-quality images and a 31% improvement on low-quality images, directly addressing ultrasound's fundamental challenge of variable image quality.

Collectively, these contributions establish the technical foundations for robust clinical imaging through efficient segmentation architectures, the integration of imaging with clinical data, explicit quality modeling, and comprehensive clinical validation across two medical imaging modalities.

Keywords: Deep Learning, Computer Vision, Biomedical engineering.

Resum

Aquesta tesi avança l'anàlisi automatitzada d'imatge mèdica introduint quatre sistemes d'aprenentatge profund que aborden barreres tècniques centrals per a la implementació clínica, incloent l'eficiència computacional, la variabilitat de qualitat i la integració robusta amb dades clíniques. A través d'un compendi d'articles, desenvolupem solucions d'última generació en ultrasons i ressonància magnètica.

En primer lloc, presentem un sistema *end-to-end* per a la mesura del gruix íntima-mitjana carotidi (CIMT) en ultrasons, aconseguint una caracterització de placa d'última generació i millorant la velocitat 20× (de 0,79 a 0,04 s/imatge). El sistema proporciona resultats complets (segmentació, mesures i detecció), eliminant requisits de post-processament específics.

En segon lloc, aprofitant característiques del nostre model *end-to-end*, som pioners en la seva integració en models de supervivència. Demostrem que els biomarcadors d'imatge apresos milloren l'estratificació del risc cardiovascular, augmentant un 20% la reclassificació del pacient respecte a les variables clíniques tradicionals.

En tercer lloc, desenvolupem una arquitectura EfficientNet-UNet++ multinivell per a la segmentació 3D de la paret carotídia en RM *black-blood* que aconsegueix un rendiment d'última generació mitjançant concatenació contextual i optimització de resolució. El sistema demostra un rendiment òptim a resolució 256×256 (6× l'original) mantenint l'eficiència via processament multinivell.

Finalment, introduïm un sistema de segmentació conscient de la qualitat amb funcions de pèrdua per modelar la qualitat durant l'entrenament. Aplicat a la paret del còlon en ultrasons, millora un 20% en imatges de qualitat mitjana i un 31% en baixa, abordant el repte de la qualitat variable.

En conjunt, aquestes contribucions estableixen fonaments tècnics per a la imatge clínica robusta mitjançant segmentació eficient, integració de dades, modelat de qualitat i validació clínica en dues modalitats.

Paraules clau: Aprenentatge profund, Visió per computador, Enginyeria biomèdica.

Acknowledgments

With great power comes great responsibility. As AI advances, those who build and apply it can choose to use it to help people —or not. When I began working with AI, I realized that its real value lay in applying it to problems where it could make a meaningful difference in people's lives. Too often, its potential is directed toward uses that do little to improve human well-being. I am grateful to the researchers who direct that power to human medicine, bringing us a step closer to easing one of humanity's greatest burdens. To the clinicians, scientists, and engineers who will use, test, and build upon the models in this thesis: this work is for you. Your scrutiny, creativity, and care are what turn ideas into better patient outcomes.

To my parents, **Leo** and **Sil**, for always believing I could go further —I hope to keep surprising you. To my siblings, **Luli** and **Tomi**, and to **Eider**, whose persistence and growth inspire me. To **Alaia**, my moral compass, for helping me see how to be better and focus my efforts on making a positive impact.

To the **MICCAI** community for taking me around the world, pushing the limits of my knowledge, and for the new friends it gave me, especially **Justin**, whose wisdom helped me both question and rediscover my confidence in academia and focus on the clinical domain. To **Marta**, whose unrelenting "What is this for?" keeps me grounded in purpose; **Miguel**, for unwavering support; **Fons**, for sharing my enthusiasm; **Martín**, whom this PhD had me run into everywhere except at home; **Sabri**, for always being there; **Fran**, for following me anywhere. To **Jesús** and **Rocío**, who helped me understand what a thesis was before I began, and lightened it with humor; **Pablo**, for teaching me ultrasound and wanting to create new things together; **Horacio**, for showing me there is no ceiling to how high one can dream; and **Barby**, my invaluable collaborator.

Finally, to my supervisors, Laura and Bea, thank you for your trust from day one. To everyone who taught me, challenged me, and will carry this work forward into clinical practice and future research —thank you.

Contents

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

During the development of this thesis, artificial intelligence (AI) has transformed from a specialized technical concept to a ubiquitous force reshaping society. In medical imaging, this transformation has been particularly profound, with deep learning methods achieving expert-level performance in tasks once thought to require years of specialized training. Yet, despite these advances, a significant gap persists between research breakthroughs and routine clinical deployment. This gap is defined by critical technical barriers: the need for computational efficiency in real-time workflows, robustness against the inherent variability of clinical data, and the challenge of integrating model outputs into complex diagnostic and prognostic frameworks.

This thesis directly confronts these barriers through a compendium of four research articles. We develop and validate a series of novel deep learning frameworks that systematically address the challenges of efficiency, data quality, and translation of image analysis into clinically meaningful data. While the primary focus is on ultrasound —a modality defined by its real-time constraints and operator-dependent quality— the principles and architectures are extended to 3D MRI to demonstrate broader applicability. Each contribution advances the state-of-the-art in a specific application while collectively building a technical foundation for clinically viable automated analysis.

1.1 The Evolution of Artificial Intelligence in Medicine

Artificial intelligence, defined as computational methods enabling machines to perform tasks requiring human intelligence, has evolved substantially within the medical domain. Modern medical AI, capable of matching or exceeding human performance on specific

diagnostic tasks, is driven by three key factors: exponential growth in computational power, the availability of large-scale medical datasets, and foundational advances in neural network architectures [1, 2]. Machine learning (ML), a subfield of AI, allows systems to learn patterns from data without being explicitly programmed. Deep learning (DL) builds upon this concept by employing multi-layered artificial neural networks to learn hierarchical representations of data. These deep models have proven particularly effective for medical image analysis, where their capacity to learn abstract features is well-suited to the complexity of visual patterns and subtle diagnostic markers. The integration of AI into clinical practice promises to address pressing healthcare challenges, such as reducing diagnostic errors, improving the efficiency of medical workflows, enabling early disease detection, and extending expert-level care to underserved populations [3, 4]. However, translating this potential into reliable clinical tools requires moving beyond simple accuracy metrics. Real-world deployment demands solutions to fundamental technical challenges in model robustness, computational efficiency, and meaningful clinical integration. This thesis is motivated by these specific challenges, aiming to develop frameworks that are not only accurate but also efficient, robust to data variability, and validated for clinical use.

1.2 Clinical Imaging Modalities and Technical Challenges

This thesis introduces novel deep learning frameworks for two of the most widely used clinical imaging modalities: ultrasound and magnetic resonance imaging (MRI). While both serve as powerful diagnostic tools, each presents distinct and significant technical challenges for automated analysis, which motivate the specific contributions of this research.

1.2.1 Modalities and Clinical Applications

Ultrasound imaging holds a unique position in clinical practice, combining interactive, real-time capabilities with the advantages of being noninvasive, radiation-free, and relatively low-cost [5]. The acquisition process itself is dynamic: a clinician uses a handheld probe to scan the patient, instantly seeing the resulting images on a screen. This allows for live, interactive examination of anatomy and function, but it also imposes the stringent requirement that any computational analysis must provide immediate feedback to be useful during the procedure. These features make it indispensable for

point-of-care diagnostics, large-scale screening, and the monitoring of chronic conditions [6]. This thesis explores two critical ultrasound applications, each presenting distinct clinical challenges that motivate the need for automated analysis.

First, we address the cardiovascular assessment of the carotid artery. The clinical objective is the early detection of atherosclerosis, the leading cause of mortality worldwide [7]. Automated measurement of the carotid intima-media thickness (CIMT) and detection of atherosclerotic plaques serve as crucial surrogate markers for assessing a patient's risk of future events like stroke and myocardial infarction [8, 9]. However, manual delineation of these fine arterial layers is a tedious process known to suffer from significant inter-observer variability, which can compromise diagnostic reliability and the statistical power of clinical trials [10]. The core computational task is therefore the precise semantic segmentation of the carotid wall layers (Figure 1.1a), from which robust and reproducible quantitative measurements can be automatically derived.

Second, we tackle the gastrointestinal monitoring of the colon wall, a key application in managing inflammatory bowel diseases (IBD) such as ulcerative colitis [11]. Transabdominal ultrasound offers a noninvasive, radiation-free method to track disease activity by measuring colon wall thickness, providing a valuable alternative to repeated invasive colonoscopies [6]. This application, however, presents an even greater technical challenge due to the extreme variability in image quality. The colon wall is a small structure whose acoustic features are often indistinguishable from surrounding tissue, and its appearance is frequently obscured by speckle noise and artifacts (Figure 1.1b). This ambiguity is so pronounced that even clinical experts exhibit high disagreement [12], highlighting the profound difficulty of the segmentation task and the urgent need for a robust automated solution.

In parallel, this work addresses the challenges of 3D black-blood MRI for high-precision cardiovascular analysis, also focusing on the segmentation of the carotid vessel wall. Unlike ultrasound, MRI offers superior soft-tissue contrast and is less operator-dependent, making it a gold standard for detailed morphological assessment of plaque burden and vessel wall remodeling. The primary imaging task is the precise segmentation of both the inner (lumen) and outer vessel wall boundaries across a 3D volume (Figure 1.1c). This allows for the direct quantification of plaque volume, a more comprehensive biomarker than CIMT. However, its clinical application is hampered by several factors. The vessel wall is a minuscule, ring-shaped structure, often less than a millimeter thick and constituting less than 0.1% of the total image volume. Accurately capturing its subtle anatomical boundaries, especially in the presence of complex plaque morphologies or near arterial bifurcations, requires a high degree of precision while

maintaining slice-to-slice spatial consistency. The time-intensive nature of manual 3D segmentation makes it impractical for routine clinical workflows, motivating the development of an efficient and accurate automated solution.

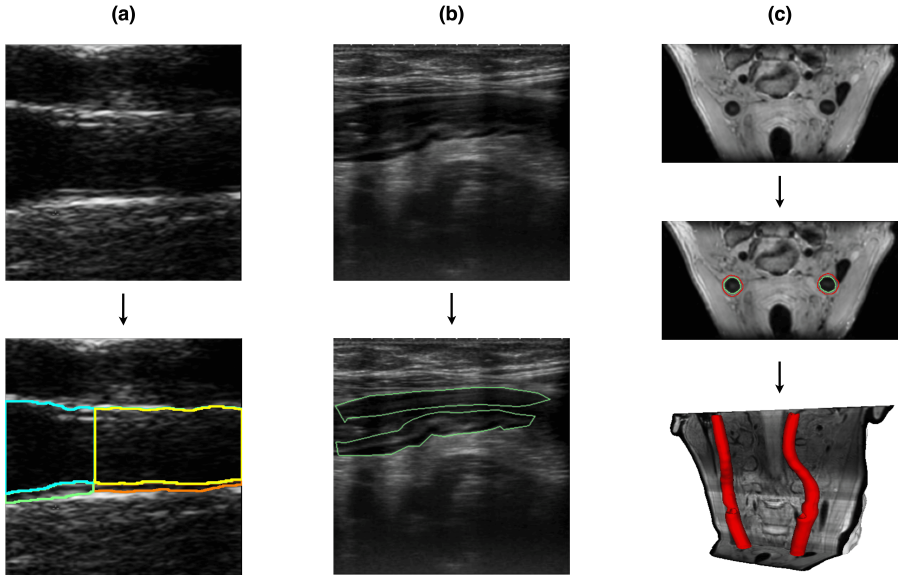

Figure 1.1: Illustrative examples of the imaging modalities and segmentation challenges addressed in this thesis. **(a)** Longitudinal B-mode ultrasound of the carotid artery, where the primary challenge is the accurate segmentation of fine wall layers. The image is part of REGICOR CCA-Seg dataset [13]. **(b)** Transabdominal ultrasound of the colon wall, highlighting the severe image quality degradation and speckle noise that complicate segmentation. The image is part of the C-TRUS dataset [12], cropped for clarity. **(c)** Axial slice from a 3D black-blood MRI of the carotid artery, showing superior soft-tissue contrast but requiring high-precision vessel wall segmentation, followed by a complete 3D segmentation. The image is part of the COSMOS dataset [14], preprocessed for clarity.

1.2.2 Core Technical Challenges

Despite their diagnostic power, the effective automation of image analysis in these modalities is hindered by several fundamental technical problems. The frameworks developed in this thesis are designed to directly address these challenges.

Robustness to Image Quality and Domain Shift

Perhaps the most significant barrier, particularly in ultrasound, is the substantial variability in image quality [15, 12]. Ultrasound images are inherently degraded by

speckle noise [16], acoustic artifacts [17], and are highly dependent on operator skill and patient anatomy [10]. This variability poses a significant challenge to model generalization, often causing models trained on a specific dataset to fail catastrophically in real-world clinical scenarios [12]. Similarly, both ultrasound and MRI models suffer from domain shift, where performance degrades when applied to data from new clinical sites, equipment, or patient populations [18]. Addressing this requires the development of architectures that can explicitly model image quality and maintain performance across diverse acquisition conditions. Crucially, this approach is contingent on two prerequisites: the availability of clinical datasets containing reliable image quality annotations, and an evaluation methodology that analyzes performance not as a single average, but stratified across these distinct quality levels.

Computational Efficiency for Clinical Workflows

The real-time, interactive nature of ultrasound imposes the most stringent constraints, requiring models that deliver feedback within milliseconds to assist the operator during the acquisition itself [5]. This is especially critical for point-of-care devices, which have limited computational power and memory. This demands architectures that are not only fast but also lightweight (i.e., having a small memory footprint), enabling them to run on affordable, portable hardware rather than requiring expensive, specialized servers [6].

While not a real-time modality, the analysis of large 3D MRI volumes also demands computational efficiency. Here, the challenge is twofold: minimizing post-processing time to avoid diagnostic bottlenecks, and ensuring the model can operate on standard clinical hardware without mandating costly GPU upgrades. The manual segmentation of a 3D volume is time-intensive, often requiring hours; automated tools must achieve comparable results within minutes. However, a model that requires a dedicated high-performance computing cluster is far less practical and scalable for widespread hospital deployment than one designed to run efficiently on existing infrastructure.

This technical need for efficiency—in terms of latency, throughput, and hardware footprint—drives the development of end-to-end frameworks that minimize computational load and avoid complex, time-consuming post-processing steps.

Precision and Integration with Clinical Decision-Making

For any AI tool to be clinically useful, it must provide precise, reliable outputs that integrate into diagnostic and prognostic workflows [6]. This presents a dual challenge.

First, there is the segmentation challenge of accurately delineating complex, low-contrast structures, which is made more difficult by the noisy annotations and high inter-observer variability common in medical imaging [12, 18]. Architectures like U-Net [19] have provided a strong foundation, but further innovation is needed for highly precise tasks. Second, the outputs of these models must extend beyond pixel-level predictions; they need to be translated into clinically meaningful biomarkers that enhance patient risk stratification, support clinical decision-making, and ultimately demonstrate improvements in patient outcomes [3, 4].

1.3 Research Objectives and Approach

The overarching goal of this thesis is to address the critical technical barriers that prevent the widespread clinical adoption of deep learning in medical image analysis. We focus on developing frameworks that are computationally efficient, clinically integrated, precise in complex anatomies, and robust to the data variability encountered in real-world practice. This is achieved by systematically tackling the challenges of efficiency, prognostic value, 3D precision, and image quality.

Research Questions

RQ1 (Chapter 2) **How can an end-to-end deep learning framework improve the efficiency and multimodal analysis of carotid artery ultrasound compared to traditional multi-stage pipelines?**

Traditional automated methods for CIMT measurement and plaque detection often depend on complex, multi-stage pipelines involving handcrafted algorithms and separate post-processing steps. This approach can be computationally slow, brittle to image variations, and difficult to optimize.

Objectives:

- To design a unified, end-to-end architecture that performs segmentation, measurement, and classification in a single forward pass.
- To eliminate the dependency on domain-specific post-processing, thereby increasing robustness and generalizability.
- To optimize the model for high computational efficiency, achieving a significant speedup (e.g., 20× to enable real-time application).

- To validate that the end-to-end approach can achieve state-of-the-art performance for atherosclerotic plaque characterization while providing multiple clinically relevant outputs.

RQ2 (Chapter 3) **Beyond the segmentation and measurement accuracy demonstrated in RQ1, does the underlying deep learning model capture latent prognostic information from carotid ultrasound images? Can this information, when integrated into survival models, significantly improve patient risk reclassification?** While clinical risk scores are the standard for prognosis, they do not exploit the complex morphological information contained within medical images. It remains an open question whether automatically learned image biomarkers can provide independent and additive predictive power for patient outcomes.

Objectives:

- To develop a methodology for extracting salient feature representations from carotid ultrasound images using a deep learning model.
- To pioneer the integration of these learned image features into an established clinical survival model (Cox proportional hazards).
- To quantitatively demonstrate that the inclusion of deep features significantly improves patient risk reclassification compared to a model using only traditional clinical variables.
- To establish a framework for bridging deep learning-based image analysis with clinical epidemiology for enhanced risk prediction.

RQ3 (Chapter 4) **How can we improve the precision of 3D carotid vessel wall segmentation in black-blood MRI while keeping computational efficiency?**

The accurate segmentation of the carotid vessel wall in 3D MRI is a significant challenge due to the structure's small size, low contrast against surrounding tissues, and complex morphological changes near plaques and bifurcations. Standard segmentation models often fail to capture fine details while maintaining spatial consistency.

Objectives:

- To design a novel architecture that performs a coarse localization followed by a targeted, high-resolution segmentation refinement.

- To incorporate a 3D contextual slice concatenation strategy, providing the model with local spatial information to improve accuracy in ambiguous regions.
- To systematically analyze the impact of input resolution on segmentation performance, identifying the optimal trade-off between detail and computational cost.
- To achieve and document state-of-the-art segmentation performance on a public 3D carotid MRI dataset.

RQ4 (Chapter 5) **Can a deep learning framework be explicitly designed to be robust to the significant image quality variability inherent in clinical ultrasound?**

A primary failure mode for AI models in clinical practice is performance degradation on low-quality images. Standard models are typically quality-agnostic, making them brittle when faced with the noise, artifacts, and operator-dependent variability common in ultrasound.

Objectives:

- To investigate the performance limitations of standard segmentation models when applied to ultrasound images of varying quality.
- To develop a novel, quality-aware framework that explicitly models image quality during the training process.
- To design a custom loss function that modulates the training objective based on image quality, forcing the model to learn robust features.
- To quantitatively demonstrate a significant performance improvement on medium- and low-quality images for a challenging segmentation task (colon wall), thereby bridging the quality gap.

1.4 Principal Contributions and Articles

The research objectives outlined in the previous section were addressed through a compendium of four articles, which together represent the principal contributions of this thesis. Each article directly corresponds to one of the research questions.

1. **End-to-End Framework for Efficient and Multimodal Carotid Analysis.** In response to RQ1, we designed and validated a unified deep learning framework

for carotid artery ultrasound analysis. This contribution demonstrates that an end-to-end approach can eliminate brittle post-processing steps, achieve a 20× speed improvement for real-time workflows, and provide a comprehensive multimodal output (segmentation, CIMT measurement, and plaque detection) while maintaining state-of-the-art accuracy.

Article: Gago, L., del Mar Vila, M., Grau, M., Remeseiro, B., & Igual, L. (2022). An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery. *Computer Methods and Programs in Biomedicine*, 223, 106954. `https://doi.org/10.1016/j.cmpb.2022.106954`
Impact factor: 6.1 (Q1 in Computer Science, Theory & Methods).

2. **Integration of Deep Features for Enhanced Cardiovascular Risk Stratification.** Addressing RQ2, this work pioneers the integration of learned imaging biomarkers into clinical survival models. Building directly upon the model developed for RQ1, we used it as a feature extractor to develop a novel framework that augments traditional clinical variables with deep features from carotid ultrasound. This resulted in a 20% improvement in patient risk reclassification and established a methodology for translating automated image analysis into clinically actionable prognostic information.

   Article: Vila, M. del M., Gago, L., Pérez-Sánchez, P., Grau, M., Remeseiro, B., & Igual, L. (2024). Deep-stratification of the cardiovascular risk by ultrasound carotid artery images. *Biomedical Signal Processing and Control*, 91, 106035. `https://doi.org/10.1016/j.bspc.2024.106035`
   Impact factor: 4.9 (Q2 in Biomedical Engineering).

3. **Context-Aware Multilevel Architecture for Precise 3D Vessel Wall Segmentation.** To answer RQ3, we developed a novel context-aware, multilevel architecture for the challenging task of 3D carotid vessel wall segmentation in MRI. The framework achieves state-of-the-art performance through a coarse-to-fine refinement strategy, the use of 3D contextual slice information, and systematic resolution optimization, showcasing a robust solution for high-precision segmentation in complex anatomies.

   Article: Gago, L., Remeseiro, B., & Igual, L. (Under Peer Review). Context-Aware Multilevel EfficientNet-UNet++ for Precise 3-D Carotid Vessel-Wall Segmentation. Submitted to *Expert Systems With Applications*.
   Impact factor: 7.5 (Q1 in Computer Science, Artificial Intelligence)

4. **Quality-Aware Framework for Robust Ultrasound Segmentation.** Confronting the challenge of data variability in RQ4, this contribution introduces a quality-aware segmentation framework. By designing a custom loss function that explicitly models image quality during training, the model learns to perform robustly across the full spectrum of clinical data. This approach yielded a 20–31% performance increase on medium-to-low quality ultrasound images, directly bridging the gap between laboratory performance and real-world clinical utility.

   Article: Gago, L., Fernández González, M. Á., Engelmann, J., Remeseiro, B., & Igual, L. (2025). Bridging the Quality Gap: Robust Colon Wall Segmentation in Noisy Transabdominal Ultrasound. *Computers in Biology and Medicine*, Volume 197, 111077. `https://doi.org/10.1016/j.compbiomed.2025.111077`
   Impact factor: 6.3 (Q1 Computer Science, Interdisciplinary Applications).

5. **Commitment to Open Science and Reproducibility.** In support of reproducible research, the source code, training pipelines, and pretrained models for the core contributions of this thesis have been made publicly available at the following repositories:

   - `https://github.com/gagolucasm/DL_CIMT_and_plaque_estimation`
   - `https://github.com/mmarvila/CA_deep_stratification`
   - `https://github.com/gagolucasm/BB-MRI-Carotid-Artery-Segmentation`
   - `https://github.com/gagolucasm/quality_gap_ultrasound_segmentation`

A complete list of the contributions derived from the doctoral work is detailed in Appendix A, including publications in journals and international conferences, as well as participation in competitive medical imaging challenges, awards, and other recognitions.

1.5 Clinical Translation and Healthcare Impact

The technical contributions of this thesis are not ends in themselves; they are designed as foundational steps toward developing clinically viable tools that address tangible problems in modern healthcare. By focusing on efficiency, predictive accuracy, and robustness, this work has the potential to translate into significant real-world impact across several key areas.

1. **Bridging the Expertise Gap in Underserved Areas.** A global shortage of trained sonographers and radiologists limits access to high-quality diagnostics.

The automated, efficient frameworks developed in this thesis (Chapters 2 and 4) are designed to function as expert assistants. In point-of-care settings, an AI tool that provides real-time, specialist-level analysis could empower general practitioners or nurses in primary care settings to perform initial screenings for cardiovascular risk or inflammatory bowel disease [6]. This democratizes expertise, reduces the need for immediate specialist consultation, and extends critical diagnostic capabilities to populations that currently lack them.

2. **Improving Diagnostic Consistency and Reliability.** A well-known limitation of ultrasound is its operator dependency, which leads to significant variability in image quality and diagnostic measurements across different clinicians and institutions [10]. This inconsistency complicates clinical trials and can lead to unreliable patient monitoring. The quality-aware framework (Chapter 5) directly confronts this problem. By delivering robust performance even on low-quality images, such a system can act as a standardizing tool, ensuring that measurements are consistent regardless of operator skill. This reliability is a prerequisite for building trust with clinicians and for gaining regulatory approval for large-scale deployment [2].

3. **Moving From Image Analysis to Patient-Specific Risk Prediction.** Modern medicine is shifting toward personalized, preventive care. This requires tools that can predict future events, not just characterize current disease. The work on integrating deep features into survival models (Chapter 3) represents a crucial step in this direction. It demonstrates that our AI models can identify subtle, sub-visual patterns in ultrasound images that correlate with a patient's long-term cardiovascular risk —patterns that are invisible to the human eye. This allows AI to augment, rather than replace, clinical judgment [3]. A clinician could use this additional prognostic information to make more informed decisions about initiating preventive treatments, such as statins, for patients who might otherwise be considered borderline risk.

4. **Streamlining the Clinical Workflow.** For any new technology to be adopted, it must fit seamlessly into existing clinical workflows without causing disruption. The extreme computational efficiency achieved in our end-to-end framework (Chapter 2) is designed for this purpose. A tool that provides immediate feedback during an ultrasound exam or reduces the analysis time for a 3D MRI from a manual task of hours to an automated one of minutes is a tool that clinicians will actually use [5]. By reducing tedious manual labor, these systems free up

valuable clinician time to focus on what matters most: patient interaction and complex decision-making.

Collectively, these advances represent a systematic progression from research prototypes toward robust clinical tools. They lay the groundwork for a future where AI-powered diagnostics are more accessible, reliable, predictive, and efficient, with the ultimate goal of improving patient outcomes in diverse clinical settings worldwide.

1.6 Thesis Structure

This thesis is structured as a compendium of four articles. Each subsequent chapter corresponds to one of the core research questions, together providing a comprehensive technical foundation for robust deep learning in medical image analysis. The chapters are organized to reflect a logical progression: from computational efficiency and clinical integration, to 3D segmentation precision and robustness against data variability.

Chapter 2: End-to-End Cardiovascular Analysis establishes the foundational end-to-end framework for efficient analysis of carotid artery ultrasound, providing multiple clinically relevant outputs. This work addresses the need for computational efficiency and the elimination of brittle post-processing steps (RQ1).

Chapter 3: Deep Features for Risk Prediction builds directly upon the feature extraction capabilities of the first contribution to pioneer the integration of learned image biomarkers into clinical survival models, demonstrating a significant enhancement in cardiovascular risk stratification (RQ2).

Chapter 4: 3D Cardiovascular Segmentation extends the scope of the thesis to high-precision 3D analysis in MRI, introducing a novel context-aware, multilevel architecture that achieves state-of-the-art performance for the challenging task of vessel wall segmentation (RQ3).

Chapter 5: Quality-Aware Segmentation addresses the fundamental challenge of image quality variability in ultrasound by developing a novel quality-aware framework. This work introduces a custom loss function that ensures robust segmentation performance even in noisy, low-quality clinical images (RQ4).

Chapter 6: Conclusions and Future Work concludes the thesis by synthesizing the main contributions across all four articles, discussing their collective impact

and limitations, and outlining future research directions toward the broad clinical deployment of robust medical imaging AI.

Chapter 2

End-to-End Cardiovascular Analysis

An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery

Lucas Gago[1], Maria del Mar Vila[1,2,3], Maria Grau[5,2,3], Beatriz Remeseiro[4*], Laura Igual[1*]

[1]Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona 08007, Spain

[2]Dept. Epidemiologia i Salut Pública, IMIM, Institut Hospital del Mar d'Investigacions Mèdiques, Dr. Aiguader 88, Barcelona 08003, Spain

[3]CIBER Enfermedades Cardiovasculares, Instituto de Salud Carlos III, Monforte de Lemos 3-5, Pabellón 11, Madrid 28029, Spain

[4]Dept. of Computer Science, Universidad de Oviedo, Campus de Gijón s/n, Gijón 33203, Spain

[5]Dept. de Medicina, Universitat de Barcelona, Carrer Casanova 143, Barcelona 08036, Spain

*These authors jointly supervised this work.

Abstract

Background and objectives: The detection and delineation of atherosclerotic plaque are usually manually performed by medical experts on the carotid artery. Evidence

suggests that this manual process is subject to errors and has a large variability between experts, equipment, and datasets. This paper proposes a robust end-to-end framework for automatic atherosclerotic plaque detection.

Methods: The proposed framework is composed of: (1) a semantic segmentation model based on U-Net, with EfficientNet as the backbone, that obtains a segmentation mask with the carotid intima-media region; and (2) a convolutional neural network designed using Bayesian optimization that simultaneously performs a regression to get the average and maximum carotid intima media thickness, and a classification to determine the presence of plaque.

Results: Our approach improves the state-of-the-art in both co and bulb territories in the REGICOR database, with more than 8000 images, while providing predictions in real-time. The correlation coefficient was 0.89 in the common carotid artery and 0.74 for bulb region, and the F1 score for atherosclerotic plaque detecting was 0.60 and 0.59, respectively. The experimentation carried out includes a comparison with other fully automatic methods for carotid intima media thickness estimation found in the literature. Additionally, we present an extensive experimental study to evaluate the robustness of our proposal, as well as its suitability and efficiency compared to different versions of the framework.

Conclusions: The proposed end-to-end framework significantly improves the automatic characterization of atherosclerotic plaque. The generation of the segmented mask can be helpful for practitioners since it allows them to evaluate and interpret the model's results by visual inspection. Furthermore, the proposed framework overcomes the limitations of previous research based on ad-hoc post-processing, which could lead to overestimations in the case of oblique forms of the carotid artery.

Keywords: Deep Learning, Semantic Segmentation, CIMT estimation, Atherosclerotic Plaque

Resum

Antecedents i objectius: La detecció i delimitació de la placa ateroscleròtica a l'artèria caròtida són realitzades, generalment, de manera manual per experts mèdics. L'evidència suggereix que aquest procés manual està subjecte a errors i presenta una gran variabilitat entre experts, equips i conjunts de dades. Aquest article proposa un marc de treball (framework) robust i d'extrem a extrem (end-to-end) per a la detecció automàtica de la placa ateroscleròtica.

Mètodes: El marc proposat es compon de: (1) un model de segmentació semàntica basat en U-Net, amb EfficientNet com a xarxa base (backbone), que obté una màscara de segmentació amb la regió íntima-mitjana de la caròtida; i (2) una xarxa neuronal convolucional dissenyada mitjançant optimització bayesiana que realitza simultàniament una regressió per obtenir el gruix íntima-mitjana carotidi (GIM) mitjà i màxim, i una classificació per determinar la presència de placa.
Resultats: El nostre enfocament millora l'estat de l'art tant en els territoris de l'artèria comuna com en els del bulb a la base de dades REGICOR, amb més de 8.000 imatges, alhora que proporciona prediccions en temps real. El coeficient de correlació va ser de 0,89 a l'artèria caròtida comuna i de 0,74 a la regió del bulb, i la puntuació F1 per a la detecció de placa ateroscleròtica va ser de 0,60 i 0,59, respectivament. L'experimentació duta a terme inclou una comparació amb altres mètodes totalment automàtics per a l'estimació del gruix íntima-mitjana carotidi trobats a la literatura. A més, presentem un extens estudi experimental per avaluar la robustesa de la nostra proposta, així com la seva idoneïtat i eficiència en comparació amb diferents versions del marc de treball.
Conclusions: El marc de treball d'extrem a extrem proposat millora significativament la caracterització automàtica de la placa ateroscleròtica. La generació de la màscara segmentada pot ser útil per als professionals mèdics, ja que els permet avaluar i interpretar els resultats del model mitjançant inspecció visual. D'altra banda, el marc proposat supera les limitacions d'investigacions anteriors basades en postprocessaments ad hoc, que podien conduir a sobreestimacions en el cas de formes obliqües de l'artèria caròtida.

Paraules clau: Aprenentatge profund, Segmentació semàntica, Estimació del GIM (Gruix Íntima-Mitjana), Placa ateroscleròtica.

2.1 Introduction

The term atherosclerosis refers to a progressive disease characterized by the accumulation of lipids and fibrous substances in the large arteries [20]. Data from several studies suggest that this process can begin in early childhood [21] and worsens with age, while it can eventually lead to reduced blood flow through the affected vessel [22]. On top of that, atherosclerosis affecting the carotid artery (CA) is considered to be the main

clinical manifestation of cardiovascular disease.

Cardiovascular disease (CVD) is the main cause of death in developed countries, and one of the leading causes of disease burden [23]. For these reasons, it is clinically essential to be able to accurately detect and mark plaque formation, thus allowing the progress of atherosclerosis to be controlled and monitored. Carotid intima-media thickness (CIMT), which estimates the width of the two deepest layers of arterial walls, is the most common sign of atherosclerosis development. CIMT and atherosclerotic plaque formation have also been shown to be a risk factor for stroke [9], coronary artery disease [24], and myocardial infarction [25]. Mean CIMT and maximum CIMT are both included in this study for their clinical interest. Firstly, according to previous studies, mean CIMT is associated with CVD risk factors [13]. Secondly, maximum CIMT is used to detect atherosclerotic plaque, which is the common basis of CVDs [8].

The procedure of detecting early atherosclerotic vascular diseases —CIMT estimation— through ultrasound imaging is a safe, non-invasive, and cost-effective method [8]. Carotid arterial wall assessment may include the common (CCA), internal, or bulb territories of the carotid artery. Measurements and monitoring of CIMT, atherosclerotic carotid plaque, and CA diameter are crucial and are regularly evaluated with high-resolution ultrasound images and videos. For the most part, CA delineations are manually performed by medical experts, but evidence suggests that they are subject to errors and have significant variability between different experts, equipment, and datasets [10]. Consequently, the availability of automatic methods for a robust and rigorous CIMT measurement and plaque delineation is highly desirable [26].

There is a growing interest in developing and implementing computer vision systems that can be integrated into real clinical practice. The most successful type of model used for computer vision tasks to date is convolutional neural networks (CNNs), which are made up of multiple layers of convolutional filters that progressively transform the input to extract some relevant features that are ultimately used to solve a learning task.

In this context, fully convolutional networks (FCNs) have proven effective in semantically segmenting different regions of the CA in ultrasound images [27]. More specifically, del Mar et al. proposed a fully automatic method based on semantic segmentation for CIMT estimation and plaque detection. The main drawback of this method is that it applies an ad-hoc post-processing procedure after the semantic segmentation step, which needs domain knowledge and thus limits its generalization ability.

Our study aims to contribute to this growing area of research by exploring fully

automatic methods based on semantic segmentation and CNNs. The objective is to define a robust end-to-end framework, without any prior knowledge and handcrafted algorithms, to assist medical practitioners in accurately determining CIMT and detecting plaque through ultrasound imaging.

2.1.1 Related Work

Among the primary techniques for CIM region delineation we can find edge detection [28], active contours [29], and snakes [30]. More recently, machine learning [31] and deep learning approaches [32, 27, 33, 34] have been proposed.

The interested reader is referred to [35] for an updated review study in which the methods are classified into three. The first-generation technologies were low-level segmentation approaches that employed traditional image processing techniques based on thresholding to get the lumen-intima and media-adventitia boundaries and then measured the mean distance using a caliper-based solution. The second generation used contour-based procedures that utilized parametric or geometric curves. Some of them were semi-automatic, requiring user interaction for initialization and/or correction of the results. In contrast, fully automatic methods do not require any user interaction, therefore being more scalable and reproducible. The third-generation models use artificial intelligence technologies such as machine learning and deep learning.

Table 2.1 presents a comparison of different fully automatic methods for CIMT estimation found in recent literature and includes some useful information such as the artery territories considered, the number of images used for evaluation purposes, and the mean CIMT error.

Table 2.1: Comparison of fully automatic methods for CIMT estimation

Article	Year	Method	Artery territory	No. of images	Mean CIMT error (mm)
Molinari et al. [28]	2012	Edge detection	CCA	365	0.078 ± 0.112
Menchón-Lara and Sancho-Gómez [32]	2014	Autoencoders for feature extraction	CCA	55	0.0499 ± 0.0489
Ikeda et al. [36]	2017	Bulb edge point detection	Bulb	649	0.0106 ± 0.0031
Qian and Yang [31]	2018	Patch-based machine learning	CCA	29	0.34 ± 0.10
Biswas et al. [33]	2018	Fully convolutional networks	CCA	396	0.126 ± 0.134
Biswas et al. [34]	2020	Patch classification and segmentation	CCA	250	0.093 ± 0.0637
del Mar et al. [27]	2020	FCN Semantic Segmentation with ad-hoc post-processing	CCA	4751	0.022 ± 0.1254
			Bulb	3733	0.06 ± 0.2749
Lian et al. [37]	2021	Poly-line estimation with DQN	CCA	4351	0.06 ± 0.04
Our proposal	2022	FCN Semantic Segmentation and CNN for CIMT estimation	CCA	4727	0.0058 ± 0.0902
			Bulb	3721	0.0096 ± 0.1791

Molinari et al. [28] proposed an automatic procedure that uses edge detection and statistical classification for CA recognition and segmentation for IMT measurement. The experimentation, carried out on 365 CCA images, showed that their method underestimated the IMT values, despite outperforming previous approaches. Also focused on CCA territory, Menchón-Lara and Sancho-Gómez [32] presented a deep

learning method based on artificial neural networks and autoencoders to identify CIMT boundaries. The proposed method was tested on a set of 55 longitudinal ultrasound images of CCA.

More recently, Ikeda et al. [36] proposed an automatic method based on carotid geometry and pixel classification to locate the bulb edges and used them for IMT measurement. The method was evaluated on 649 images with different types of plaque and image resolutions, achieving a coefficient of correlation of 0.998.

On the other hand, Qian and Yang [31] proposed a model that integrates random forest and an auto-context model for pixel-wise classification, reporting a dice similarity coefficient of 0.81 in only 29 ultrasound images.

Biswas et al. [33] presented a two-stage deep learning system composed of a convolution layer-based encoder for feature extraction and a fully convolutional network-based decoder for image segmentation. Further work by Biswas et al. [34] includes a segmentation procedure on image patches previously classified.

For their part, del Mar et al. [27] proposed a single-step approach for automatic CA image interpretation, able to be trained in both CCA and bulb territories. This method was composed of a semantic segmentation model and an ad-hoc post-processing module for CIMT estimation. It was tested on REGICOR database [13], which is composed of more than 8000 images. Their results reach a correlation coefficient of 0.81 in CIMT estimation, and a CIMT mean error of 0.02 and 0.06mm in CCA and bulb images, respectively. Performing a segmentation of the ultrasound image before CIMT estimation showed promising results. However, the ad-hoc post-processing applied to the segmentation mask, which includes predefined morphological operations and cropping off part of the image, may limit the robustness of the proposal.

Focusing on automatic estimation but with less amount of data, Lian et al. [37] used a deep Q-network to adjust 15 points to the near-wall, far-wall, and intima-lumen interface, to later adjust poly-lines. In addition, in the reward calculation step, they incorporated anatomical priors related to straightness and parallelism, increasing the performance but limiting the model's ability to detect clinically significant outliers.

Finally, it is worth noting that some of these studies [28, 27] compute CIMT by dividing the CIM segmentation mask into vertical regions, which could lead to over-estimations in case of oblique forms of the CA [38].

2.1.2 Contributions

In this paper, we present an end-to-end framework for CIMT estimation and atherosclerotic plaque detection. The proposed framework is, in fact, a fully automatic system comprised of two modules: (1) a FCN for semantic segmentation and (2) a CNN for classification and regression. The segmentation model is composed of a light architecture with a pre-trained feature extractor as its backbone and aggressive data augmentation. The CNN takes as input both the original image and the mask provided by the segmentation module and generates a prediction of the average CIMT, the maximum CIMT, and the presence of atherosclerotic plaque.

The main contributions of this research are the following:

1. An end-to-end framework composed of two modules: (1) a semantic segmentation model, based on U-Net with EfficientNet as the backbone, achieving state-of-the-art results in CCA and bulb territories; and (2) a regression and classification model, based on a CNN designed using Bayesian optimization, capable of making real-time predictions of the maximum and average CIMT and the presence of atherosclerotic plaque in CCA and bulb territories.

2. An evaluation of the robustness of the proposed framework as well as its suitability and efficiency compared to different versions of it. In particular, a comparative study showing that training a model with a joint feature extractor for CIMT and plaque values can lead to better results than using three independent models.

3. A comparison with other fully automatic methods for CIMT estimation found in the literature, including a detailed comparison with a previous state-of-the-art approach using REGICOR database, with more than 8000 images corresponding to CCA and bulb territories.

The rest of the paper is structured as follows. Section 2.2 details the proposed end-to-end framework for CIMT estimation and plaque detection. Section 2.3 introduces the datasets used for evaluation purposes and the design of the experiments performed, followed by the results achieved. Finally, Section 2.4 closes the paper with the conclusions and future challenges.

2.2 Methodology

We propose a fully automatic framework that predicts both maximum and average CIMT, and the presence of atherosclerotic plaque without any domain knowledge

or metadata from the image. Figure 2.1 depicts the architecture of the proposed framework, which is comprised of a semantic segmentation network followed by a CNN that performs classification and regression tasks. A single CNN is used to simultaneously predict three target values, assuming this could lead to a better feature extractor and better performance on the plaque detection problem (see Section 2.3.4 for the validation of this assumption). The input size is 445x470 pixels, which corresponds to the original resolution of the REGICOR database. The two modules of the framework are subsequently described in depth.

2.2.1 Semantic Segmentation

Carotid artery semantic segmentation consists of classifying each pixel of the input ultrasound image as one of lumen, far wall, near wall, CIM, bulb, and CIM-bulb region. Figure 2.2 shows an example for each territory, with a legend that details the segmentation labels. The semantic segmentation changes the representation of the image into something more meaningful and easy to analyze.

A U-Net [39] was used to perform the carotid artery semantic segmentation. U-Net is an asymmetrical segmentation network that presents skip connections between down-sampling and up-sampling paths to improve the quality of the segmentation mask by providing local information to the encoded global information in the up-sampling process. The network is composed of three parts: the down-sampling, bottleneck, and up-sampling. Down-sampling is consists of four blocks containing 3×3 convolutional layers with batch normalization [40] followed by 2×2 max-pooling layers. At the end of each block, a skip connection is sent to the symmetric up-sampling module. The bottleneck is built from two convolutional layers with batch normalization and dropout [41] to reduce the overfitting. The up-sampling path also consists of four blocks, made up of transposed convolutions with stride 2, a concatenation with the corresponding feature map from down-sampling (skip connection), and 3×3 convolutional layers with batch normalization.

In our proposed architecture, we use EfficientNet B0 [42] as a lightweight feature extractor, pre-trained on ImageNet [43]. In particular, we replaced the down-sampling component of the U-Net with a pre-trained EfficientNet B0, while the bottleneck and up-sampling maintain the original U-Net architecture. Skip connections are sent from the first, second, third, and fifth blocks of EfficientNet B0, while the output is connected to the bottleneck part of U-Net.

The EfficientNet family of networks was generated using neural architecture search

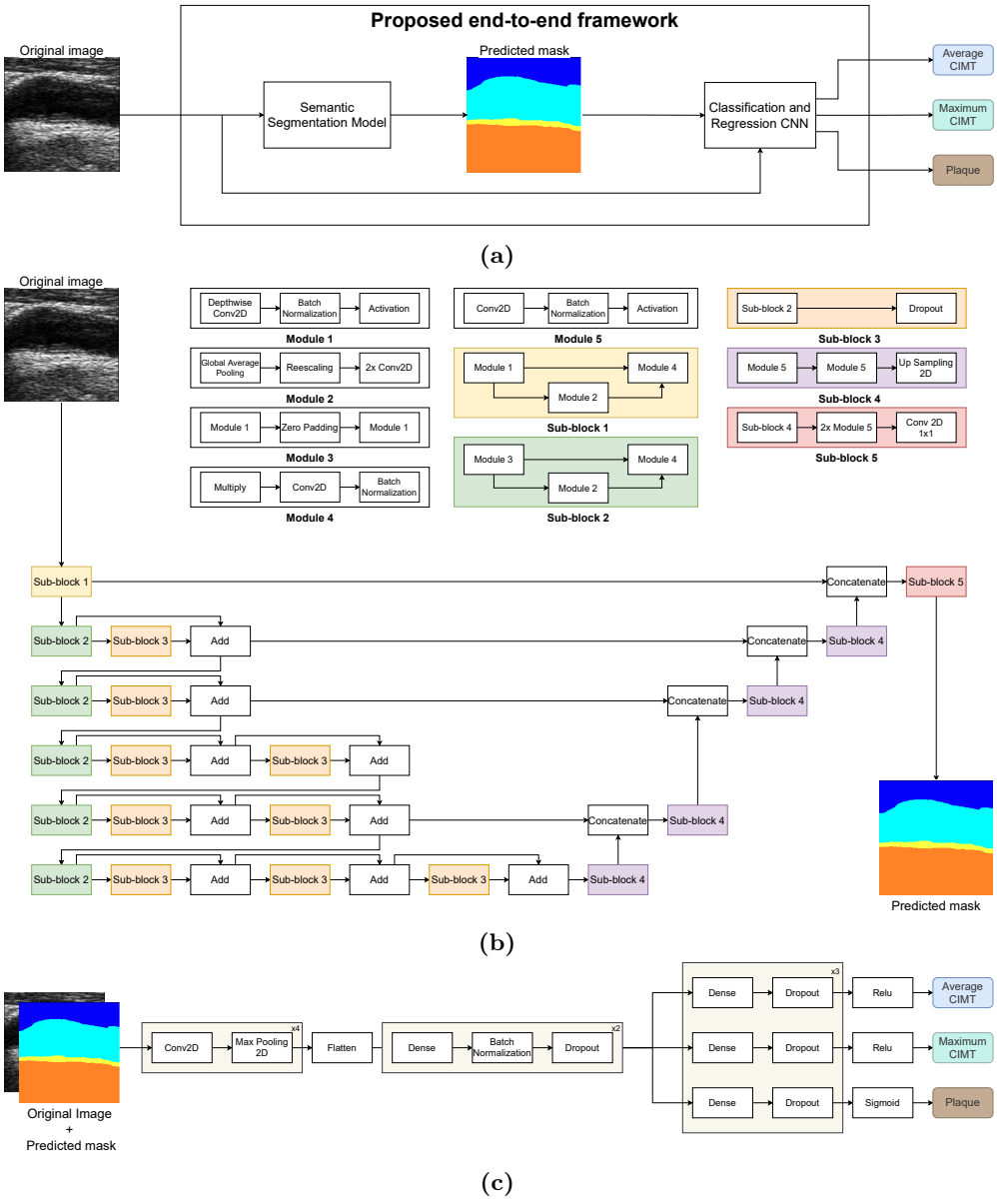

Figure 2.1: (a) The proposed end-to-end framework composed of two modules: (b) the semantic segmentation model, and (c) the classification and regression CNN model architecture defined by Bayesian optimization.

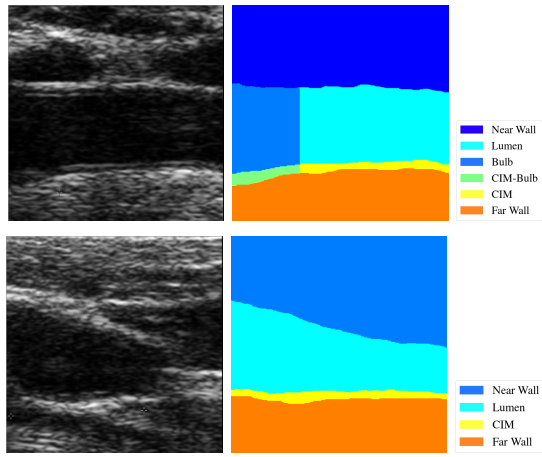

Figure 2.2: Ultrasound images and their corresponding segmentation masks from the ground truth. Top row: CCA territory (six labels). Bottom row: Bulb (four labels).

and has been proven capable of achieving high accuracy despite being much smaller and faster than previous models. EfficientNet B0 is the smallest architecture of the ones proposed, achieving 93.5% top-5 accuracy on the ImageNet validation set with only 5.3 million parameters. The EfficientNet networks are composed of different combinations of the sub-blocks from one to five, defined in Figure 2.1(a).

Light architectures such as EfficientNet B0 helps to avoid overfitting and reduce training times. Moreover, data augmentation techniques are useful to prevent the overfitting caused by training models with a low number of samples. For this reason, we applied several data augmentation techniques, including scale transformations, elastic transformations, perspective transformations, affine transformations, JPEG compression, speckle noise, motion blur, variation of hue and saturation, histogram equalization, cropping, and padding. The motivation behind the design of the data augmentation pipeline was to improve the generalization capacity of the system and make it more robust to ultrasound imaging artifacts like speckle [44].

Regarding the loss function of the semantic segmentation network, we defined a custom loss (Loss) composed of the Dice coefficient (DC) and the focal loss (FL) [45], defined as follows:

$$\text{Loss} = \text{DC} + \text{FL} \tag{2.1}$$

$$\text{DC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.2}$$

$$\mathrm{FL}\left(p_{\mathrm{t}}\right) = -\alpha_{\mathrm{t}}\left(1-p_{\mathrm{t}}\right)^{\gamma}\log\left(p_{\mathrm{t}}\right) \tag{2.3}$$

$$p_{\mathrm{t}} = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{2.4}$$

where TP, FP, and FN stand for true positives, false positives, and false negatives, respectively; $p \in [0, 1]$ is the model's estimated probability for the class with label y, p_{t} is defined for notational convenience, α is a weighting factor, and γ is the focusing parameter Notice that the focal loss component has been proven useful in segmentation problems with under-represented classes [45], as is our case with the CIM region class significantly less represented than the other classes.

2.2.2 CIMT estimation and plaque detection

We propose a regression and classification CNN model to predict average and maximum CIMT values as well as the presence of plaque in ultrasound images. Previous research relies on ad-hoc post-processing of the segmentation results [27], limiting the ability of the system to generalize. Keeping that in mind, our motivation is to eliminate any prior knowledge and handcrafted algorithms from the pipeline, making it fully automatic and able to perform well without depending on the origin, scale, or quality of the input data. The combination between the semantic segmentation model and CNN for CIMT estimation and plaque detection (see Figure 2.1) allows us to have an end-to-end fully automatic framework that can be trained without any domain-specific knowledge given tagged examples.

The CNN input is the concatenation of the original image and the segmentation mask provided by the segmentation model previously applied, with the idea of maximizing the information available to the model when predicting the presence of atherosclerotic plaque. Regarding the CNN architecture, it was tuned using a Bayesian optimization [46] given some design constraints. The system selects a variable number of convolutional layers followed by a max-pooling operation. Then, the last layer is flattened and the optimization selects a series of configurable dense layers followed by batch normalization and dropout, with a rate of 25%. At this point, the network is divided into three branches to predict one target value per branch (average CIMT, maximum CIMT, plaque detection), made up of a variable number of blocks of dense layers followed by dropout (see Figure 2.1c).

The activation function used for average and maximum CIMT estimation is the

rectified linear unit (ReLU) [47], and the sigmoid for plaque detection. As for the loss functions, we used the mean squared error (MSE) for average and maximum CIMT and binary cross-entropy (BCE) for plaque detection, defined as:

$$\text{MSE} = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{2.5}$$

$$\text{BCE} = -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \tag{2.6}$$

where y_i is the truth value, and $\hat{y}_i$ is the predicted value.

2.3 Experimental study

This section introduces datasets used for evaluation purposes and implementation details of the proposed framework. Moreover, experiments carried out and the results achieved are presented and discussed, including a comparison with the state-of-the-art in REGICOR dataset.

2.3.1 Datasets

The proposed framework was evaluated on the REGICOR database, one of the largest image collections available for the problem at hand [27]. REGICOR consists of a sample of 2379 subjects from Girona's Heart Registry [13]. Images were collected from 2007 to 2010, and the subjects represent general population aged 35 to 84. Two trained sonographers performed the CA ultrasound (US) scans with an Acuson XP128 US system equipped with an L75-10 MHz transducer and a computer program extended frequency (Siemens-Acuson). US longitudinal images were obtained in B-mode with a resolution of 23.5 pixels/mm. Original images were saved in DICOM format and then converted to PNG. The set of images collected for each patient was obtained from left and right CA in two different territories (CCA and bulb), resulting in a total of 8448 images (4727 CCA images, and 3721 bulb images). CIMT reference values, given by the Amsterdam Medical Center, were used as the ground-truth (GT) for CIMT estimation. Regarding the GT for plaque detection, it was obtained using the provided CIMT reference values and applying the Mannheim consensus [8]. Furthermore, images containing plaque were finally supervised by an expert.

Besides the GT for CIMT estimation and plaque detection, a segmentation GT was defined for a subset of REGICOR images [27]. In order to obtain it, an expert manually

delineated and labeled different regions of original images, using six labels for CCA and four for the bulb (see Figure 2.2). Only a representative subset of REGICOR images was labeled, including 159 CCA images (50 with plaque and 109 without plaque), and 172 bulb images (68 with plaque and 104 without). These labeled subsets will be referred to as REGICOR CCA-Seg and REGICOR Bulb-Seg from now on, respectively. Table 2.2 presents a summary of the datasets and their main characteristics.

Table 2.2: Summary of the different datasets of US images used in this research

Dataset	Territory	Information available	No. images	Images with plaque
REGICOR CCA-Seg	CCA	Manually segmented masks	159	50 (31.47%)
REGICOR Bulb-Seg	Bulb	Manually segmented masks	172	68 (39.53%)
REGICOR CCA	CCA	Avg. and max. CIMT and plaque	4727	50 (1.06%)
REGICOR Bulb	Bulb	Avg. and max. CIMT and plaque	3721	262 (7.04%)

2.3.2 Implementation details

Our proposed framework is implemented on Tensorflow [48], and the code will be publicly available after paper acceptance[1]. It is made up of two modules, the semantic segmentation model and the regression and classification CNN, the implementation of which is given below.

Semantic Segmentation Model

For the semantic segmentation module, we used a U-Net network with EfficientNet pre-trained on ImageNet as the backbone (see Section 2.2.1). The model is composed of 10.1 million parameters and was trained to segment all available classes, even though we are mainly interested in CIMT region. The reason is that models achieve better performance in CIMT class after being trained with all the classes, as demonstrated in [27]. For the custom loss, the weighting factor α was set to 0.25 and the focusing parameter γ to 2.0. Aggressive data augmentation was performed, ensuring that the image scale is not altered or no transformation is performed that could lead to information loss.

We used Adam [49] as the optimizer, a gradient descent method that is based on adaptive estimation of first-order and second-order moments. The parameters β_1 and β_2, representing the exponential decay rate for the first and second-moment estimates, were set to 0.9 and 0.999, respectively.

[1] `https://github.com/gagolucasm/DL_CIMT_and_plaque_estimation`

Note that early stopping was used with a patience of 20, and the learning rate was reduced on plateau with a patience of 10 and a factor of 0.1. The initial learning rate was 1e−4. The models were trained with full resolution images and a batch size of 4.

Regression and Classification CNN

The regression and classification CNN model defined in Section 2.2.2 takes the mask predicted by the segmentation network and the original image as input and uses them to predict three outputs per image: average CIMT, maximum CIMT, and plaque. Notice that the three outputs share the feature extraction part of the CNN, which was designed using a Bayesian optimization tuned with a Gaussian process trained on REGICOR CCA dataset. The motivation here is that adding more information could lead to an improvement in performance, offsetting the additional calculation.

In this case, we also used Adam as the optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The loss for the CNN was weighted, focusing more on the average and maximum CIMT outputs. This was motivated by the convergence problems for plaque prediction in REGICOR CCA dataset due to the class imbalance. Weights were set to 0.4 for average and maximum CIMT outputs, and 0.2 for plaque. In the plaque classification problem, classes were not weighted because, in our preliminary experiments, the results did not improve.

Early stopping was used with a patience of 50, and the learning rate was reduced on plateau with a patience of 15 and a factor of 0.1. Initial learning rate was 1e−3. The models were trained with full resolution images and a batch size of 16. It should be noted that mixed-precision was used to accelerate the training process and reduce the memory footprint.

2.3.3 Performance measures

For the evaluation of the semantic segmentation module, we computed a standard metric in this type of problem: the intersection over union (IoU). This metric measures the number of pixels in common between the target and prediction segmentation masks divided by the total number of pixels present across both masks:

$$\text{IoU} = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}} \tag{2.7}$$

Precision and sensitivity were also used as performance metrics:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.9}$$

where TP, FP, and FN stand for true positives, false positives, and false negatives, respectively. Notice that, for evaluation purposes, the CIM region is considered positive and the background (i.e., the combination of all other classes) negative, since the CIM region is our main focus.

To evaluate the performance of the method in predicting CIMT, we used the Pearson correlation coefficient (CC), the mean absolute error (MAE), and the mean squared error (MSE). Pearson correlation coefficient is defined as:

$$\text{CC}_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.10}$$

where X and Y are a pair of random variables, cov is the covariance, and σ_X and σ_Y are standard deviations of X and Y, respectively.

The MAE is defined as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{2.11}$$

where y_i is the truth value and $\hat{y}_i$ is the prediction. See Eq. (2.5) for the definition of the MSE.

For the plaque classification task, in addition to the sensitivity defined in Eq. (2.8), we also computed the accuracy, specificity, and F1 Score:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.12}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.13}$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN))} \tag{2.14}$$

Note that, for these measures, the presence of plaque is considered positive and its absence, negative.

2.3.4 Experimental Results

Three experiments were defined to evaluate the robustness of the proposed end-to-end network.

- Experiment 1: Evaluation of the semantic segmentation module. The objective is to measure the performance of the proposed module to segment different regions of CA and to analyze its suitability for plaque detection in comparison with previous research.

- Experiment 2: Evaluation of the impact of input data in the regression and classification CNN. The objective is to analyze whether the additional information and computation of the semantic segmentation model are necessary, and the combined use of the predicted segmentation mask and the original image as input data.

- Experiment 3: Evaluation of one versus three CNNs for plaque detection. The objective is to gain insight into its effect on the overall performance of training a single model or three independent models to predict the three target values (average CIMT, maximum CIMT, and plaque).

Table 2.3 summarizes the three experiments carried out, including datasets and performance measures used. Note that all the models were trained with an NVIDIA GeForce RTX 2080ti 11GB GPU, and the code to reproduce all the experiments will be publicly available with the implementation of the framework (see Section 2.3.2).

Table 2.3: Summary of the experiments designed for evaluation purposes

Experiments	Datasets	Train split	Performance metrics
1) Evaluation the of semantic segmentation module	REGICOR CCA-Seg REGICOR Bulb-Seg	90% train 10% test	DICE, IoU, Precision, Sensitivity
	REGICOR CCA REGICOR Bulb	100% test	Accuracy, Sensitivity, Specificity, F1 Score
2) Evaluation of the impact of input data in the regression and classification CNN	REGICOR CCA REGICOR Bulb	60% train 20% val 20% test	MAE, MSE, CC, Accuracy, Sensitivity, Specificity, F1 Score
3) Evaluation of one versus three CNNs for plaque detection	REGICOR CCA REGICOR Bulb	60% train 20% val 20% test	Accuracy, Sensitivity, Specificity, F1 Score

Experiment 1: Evaluation of the semantic segmentation module

Experiment 1 was divided into two parts: (1) a quantitative and qualitative analysis of the segmentation results on the two databases considered (REGICOR CCA-Seg and

Bulb-Seg), and (2) a comparison of results for CIMT estimation and plaque detection on REGICOR CCA and Bulb full datasets using the post-processing proposed in [27].

Table 2.4 shows the results obtained with the two methods, the segmentation network used in [27] and our segmentation module, on both datasets (REGICOR CCA-Seg and Bulb-Seg). As can be observed, our model outperforms the one proposed by del Mar et al. [27], regardless of the metric considered. More specifically, our model performs better on the CCA-Seg dataset, but there is a more significant improvement over previous results on the Bulb-Seg dataset. In this sense, it is worth mentioning that image quality in the bulb region is lower, with poorer contrast and more affected by noise.

Table 2.4: Results for semantic segmentation on REGICOR CCA-Seg and Bulb-Seg test sets. Only the CIMT region and background are considered. The best results per dataset are marked in bold face.

Dataset	Method	IoU	Dice	Precision	Sensitivity
REGICOR CCA-Seg	del Mar et al. [27] seg.	0.7121	0.8299	0.8217	0.8530
	Our seg. proposal	**0.8712**	**0.9959**	**0.9976**	**0.9942**
REGICOR Bulb-Seg	del Mar et al. [27] seg.	0.5711	0.6945	0.6930	0.7313
	Our seg. proposal	**0.9273**	**0.9694**	**0.9741**	**0.9721**

Figure 2.3 depicts some representative examples of predictions obtained with the two methods, for a qualitative comparison. As can be seen, our proposal shows better region connectivity and does not generate erroneous isolated regions. Prediction time is 0.026 seconds, an order of magnitude faster than [27] (0.79s) due to the smaller network size. Note that time per frame was measured using a GeForce Titan X (Pascal) 12GB GPU from NVIDIA in [27], while we used a GeForce RTX 2080ti 11GB GPU also from NVIDIA.

Semantic segmentation models provide a segmentation mask categorizing each pixel into a class. Based on this information, del Mar et al. [27] proposed an ad-hoc post-processing procedure for CIMT estimation. In this experiment, we applied the same procedure to the results obtained with our semantic segmentation model to measure their impact on plaque detection. The post-processing procedure is based on morphological operators and prior domain knowledge and is detailed in [27].

Table 2.5 shows the results for plaque detection compared with the most competitive ones reported so far for REGICOR datasets [27]. It's important to mention that the class plaque is underrepresented in both datasets, with 1.06% of total images in REGICOR CCA and 7.04% in REGICOR Bulb. As can be seen, the results obtained

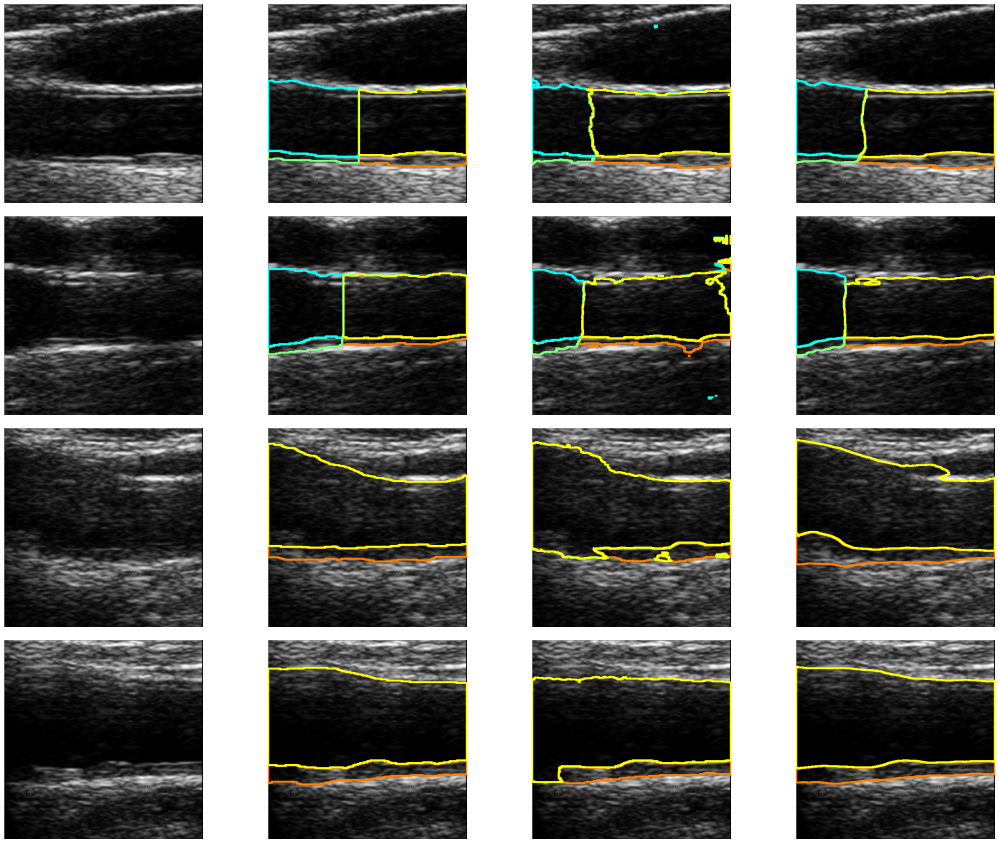

Figure 2.3: Qualitative results: (a) Four original images, (b) the ground truth, (c) the segmentation results obtained in [27], and (d) the segmentation results achieved with our proposed method. The first two rows correspond to the REGICOR CCA-Seg, whilst the last two are from REGICOR Bulb-Seg.

with the proposed method are consistently better than those obtained in [27] regardless of the metric applied, thus confirming the superior performance of our proposal not only in terms of segmentation but also in plaque detection.

Table 2.5: Results for plaque detection on REGICOR CCA and Bulb full datasets using the post-processing proposed in [27]. The best results per dataset are marked in bold face.

Dataset	Method	Plaques/total images	Accuracy	Sensitivity	Specificity	F1 Score
REGICOR CCA	del Mar et al. [27] seg.	50/4727	0.9645	0.8000	0.9663	0.3226
	Our seg. proposal	50/4727	**0.9725**	**0.8800**	**0.9735**	**0.4037**
REGICOR Bulb	del Mar et al. [27] seg.	262/3721	0.7809	0.7832	0.7500	0.3262
	Our seg. proposal	262/3721	**0.8126**	**0.9014**	**0.8054**	**0.4197**

Experiment 2: Evaluation of the impact of input data in the regression and classification CNN

For the second experiment, we conducted an ablation study in which we eliminate the segmentation module from our input to determine the contribution of the component to the overall system. The regression and classification CNN was fed with different input data: (1) only the original image, (2) only the predicted segmentation mask, or (3) the concatenation of both (our proposal). Achieving similar results with only the original image as input data would mean that we could further reduce the complexity of the system, with the disadvantage of decreasing the interpretability of the model.

Table 2.6 shows the results obtained in this experiment for maximum and average CIMT values. Here, we can appreciate a clear improvement over previous work [27] in every metric and for both datasets, all with a p-value of < 0.001. Using both the original image and the segmentation mask, our proposal, translates into a consistent improvement of results with respect to CNN versions from segmentation mask only and from original image only. Table 2.6 contains information about the mean average error (MAE), mean squared error (MSE), and correlation coefficient (CC) for the maximum and average CIMT prediction. This model performs significantly better in CCA than in bulb, where the quality of the segmentation is lower. In CCA, it obtains best results in every measured metric, showing a clear use of the additional information provided in the input. In REGICOR Bulb, the CNN from segmentation mask performs best at the maximum estimate of CIMT, while our proposal outperforms in predicting average CIMT.

Results obtained when applying the post-processing presented in [27] to our segmentation masks (Experiment 1) can be considered a more demanding baseline and

are still lower on each metric, indicating that the performance increase cannot be explained solely by new segmentations. There seems to be a positive impact from using a CNN for regression and classification. It is worth noting that CNN architecture was optimized for REGICOR CCA, and this could show penalizing results in REGICOR Bulb.

Table 2.7 shows the results for plaque classification in the test set for all the experiments. Our proposal performs significantly better on REGICOR Bulb, achieving an F1 score of 0.5909, indicating a balance between sensitivity and precision not exhibited by other experiments. On REGICOR CCA, it keeps the sensitivity level of both the Vila et al. [27] proposal and [27] post-processing applied to our segmentation masks, but with a significant increase in precision, reflected in the F1 score of 0.6000, up from 0.2667 and 0.3158 from these previous experiments. Using only the original image as input results in a very low sensitivity in the plaque prediction problem, meaning that the system cannot correctly predict images with high CIMT values, underestimating the plaque prediction. Using the segmentation mask as the only input achieves a lower performance than our proposal, which indicates that the additional information provided by the concatenation of the original image and segmentation mask has a positive impact on the plaque detection problem. In REGICOR CCA, while the F1 score is 0.6000 using only the segmentation mask and using the concatenation of the segmentation mask and the original image, the sensitivity of the latter is higher, which is 0.8571 compared to 0.5714.

Figure 2.4 includes a representation of Bland-Altman and scatter plots for the results of our proposal on the test sets of both REGICOR CCA and REGICOR Bulb. The model performs well in REGICOR CCA, with a mean error of -0.02mm. The variability seems to be consistent without any appreciable trend between CIMT values of 0.7 and 0.9mm. Predictions of average CIMT in images with values between 0.45 and 0.7mm are invariably near 0.55mm, indicating the model has difficulties in analyzing cases in this range. As for REGICOR Bulb, there is a trend towards underestimating CIMT as it gets higher, showing difficulties in the prediction of outliers. The mean error is -0.04mm, indicating a slight underestimation.

The average processing time of this CNN block is 0.014 seconds, adding to 0.040 seconds if the segmentation step is considered; that is, the processing time is almost 20 times faster than in the previous work [27].

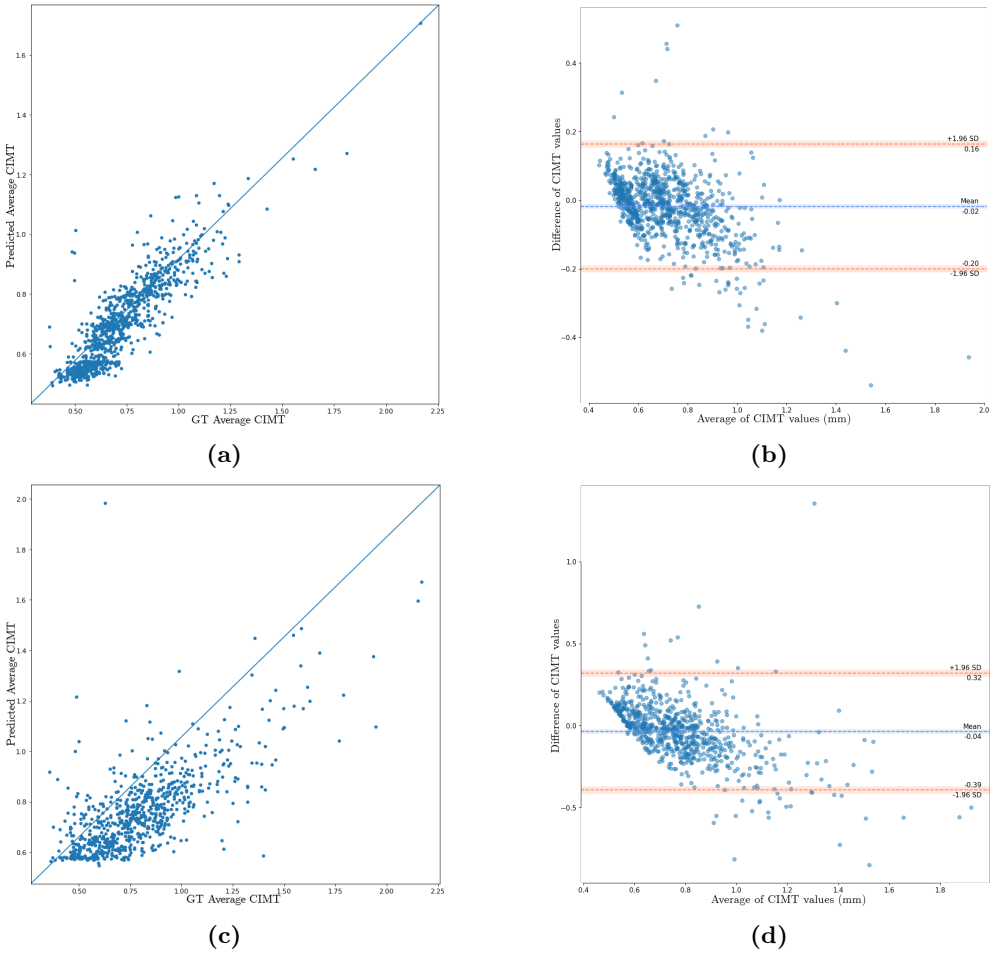

Figure 2.4: Results of average CIMT predictions obtained with our proposal (Experiment 3). (a) Correlation between average CIMT values, and (b) Bland–Altman analysis for the predicted average CIMT values. Note that the top row corresponds to the REGICOR CCA dataset, whilst the bottom row is for REGICOR Bulb.

Table 2.6: Results for maximum and average CIMT on REGICOR CCA and Bulb tests sets. The best results per dataset are marked in bold face.

				Maximum CIMT			Average CIMT		
Dataset	Exp.	Input data	Method	MAE (mm)	MSE (mm^2)	CC	MAE (mm)	MSE (mm^2)	CC
REGICOR	-	Seg. mask [27]	del Mar et al. [27] post-processing	0.1539	0.0633	0.6213	0.0906	0.0170	0.8006
CCA	1	Our seg. mask	Vila et al. [27] post-processing	0.1330	0.0477	0.6960	0.0824	0.0159	0.8312
(test set)	2	Our seg. mask	One single CNN	0.0921	0.0164	0.8367	0.0665	0.0089	0.8820
	2	Original image	One single CNN	0.1105	0.0266	0.6896	0.0818	0.0162	0.7519
	2	Orig. image + Our seg. mask	One single CNN *	**0.0896**	**0.0148**	**0.8431**	**0.0659**	**0.0082**	**0.8870**
REGICOR	-	Seg. mask [27]	del Mar et al. [27] post-processing	0.4460	0.4252	0.2673	0.1936	0.0806	0.3899
Bulb	1	Our seg. mask	Vila et al. [27] post-processing	0.3700	0.2833	0.5071	0.1901	0.0650	0.6111
(test set)	2	Our seg. mask	One single CNN	**0.1646**	**0.0589**	**0.7225**	0.1328	0.0335	0.7279
	2	Original image	One single CNN	0.2267	0.1038	0.4048	0.1781	0.0568	0.4056
	2	Orig. image + Our seg. mask	One single CNN *	0.1669	0.0595	0.7228	**0.1311**	**0.0321**	**0.7362**

* Our proposal

Table 2.7: Results for atherosclerotic plaque detection on REGICOR CCA and Bulb test sets. The best results per dataset are marked in bold face.

Dataset	Exp.	Input data	Method	Accuracy	Sensitivity	Specificity	F1 Score
REGICOR CCA (test set)	-	Seg. mask [27]	del Mar et al. [27] post-processing	0.9651	**0.8571**	0.9659	0.2667
	1	Our seg. mask	Vila et al. [27] post-processing	0.9725	**0.8571**	0.9733	0.3158
	2	Our seg. mask	One single CNN	0.9926	0.5714	0.9957	0.5333
	2	Original image	One single CNN	0.9884	0.0000	0.9957	0.0000
	2	Orig. image + Our seg. mask	One single CNN *	0.9915	**0.8571**	0.9925	**0.6000**
	3	Our seg. mask	Three CNNs (only plaque output)	**0.9947**	0.5714	0.9979	**0.6000**
	3	Original image	Three CNNs (only plaque output)	0.9926	0.0000	**1.0000**	0.0000
	3	Orig. image + Our seg. mask	Three CNNs (only plaque output)	0.9905	0.5714	0.9936	0.4706
REGICOR Bulb (test set)	-	Seg. mask [27]	del Mar et al. [27] post-processing	0.7903	0.7551	0.7928	0.3217
	1	Our seg. mask	Vila et al. [27] post-processing	0.8427	**0.9184**	0.8374	0.4348
	2	Our seg. mask	One single CNN	0.9476	0.4898	0.9799	0.5517
	2	Original image	One single CNN	0.9301	0.2245	0.9799	0.2973
	2	Orig. image + Our seg. mask	One single CNN *	**0.9516**	0.5306	0.9813	**0.5909**
	3	Our seg. mask	Three CNNs (only plaque output)	0.9422	0.3673	0.9827	0.4557
	3	Original image	Three CNNs (only plaque output)	0.9368	0.0408	**1.0000**	0.0784
	3	Orig. image + Our seg. mask	Three CNNs (only plaque output)	0.9341	0.0000	**1.0000**	0.0000

* Our proposal

Experiment 3: Evaluation of one versus three CNNs for plaque detection

We conducted a final experiment to assess whether using three CNN models, one for each target value, offers an advantage over a single CNN for the three target values (our proposal). CNN architectures used for this experiment are the same as we used in our proposal, except having one output line instead of three (see Figure 2.5). As for the training settings, they are the same as in Experiment 2.

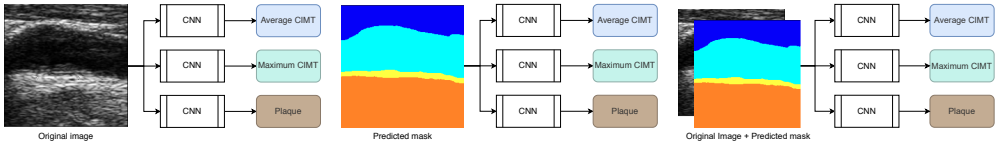

Figure 2.5: Models tested to analyze if training a single CNN with multiple outputs leads to better detection of plaque than using three individual CNNs (Experiment 3). Different input data for the CNNs: (a) the original image, (b) the segmentation mask obtained with our proposed model, and (c) a concatenation of the original image and the predicted segmentation mask. Note that the CNN blocks are the same as in our proposal, but with one single output.

Table 2.7 reports the results obtained from plaque classification using three independent CNNs, denoted as "only plaque output". Results suggest that our proposal provides better results than using three individual CNNs. On REGICOR CCA, the models trained only on plaque and only to predict the maximum CIMT cannot classify any image in the plaque category, whereas if they are trained together, they can. The same situation occurs in REGICOR Bulb dataset: the results of our model with multiple outputs and one feature extractor are better for plaque detection than using

three independent CNNs. The impossibility of detecting the presence of atherosclerotic plaque with this architecture, as well as the need to train and predict three independent models, make this solution a poor option for the problem at hand, thus demonstrating the suitability of our proposed end-to-end framework.

2.4 Conclusions and future work

The intima-media thickness and the presence of atherosclerotic plaque in the carotid artery are the most common signs of cardiovascular disease development. In this context, we present an end-to-end framework to predict average CIMT, maximum CIMT, and presence of plaque on ultrasound images of two different carotid artery territories (CCA and bulb). Our approach is composed of a semantic segmentation module to anatomically segment the input image, followed by a CNN for regression and classification of three target values (average CIMT, maximum CIMT, and plaque) fed with the original image and the predicted segmentation mask. This approach can be useful for practitioners since it allows them to evaluate and interpret the results of the model by visually inspecting the predicted segmentation masks. Moreover, the method is able to estimate CIMT in a fast and useful manner for large image datasets and enables us to eliminate the inter-observer variability usually associated with manual CIMT estimation. The proposed framework achieves state-of-the-art results in REGICOR database and reduces prediction time from 0.79 to 0.04 seconds per image, with a processing speed of 25 frames per second. We compared the semantic segmentation model with previous work, qualitatively and quantitatively, demonstrating more accurate results. Moreover, our experiments also confirm the improvement in terms of CIMT prediction and atherosclerotic plaque detection on 8290 images. Furthermore, the proposed framework overcomes the limitations of previous research [27], based on ad-hoc post-processing that computed CIMT by dividing the mask into vertical regions, which could lead to over-estimations in case of oblique forms of the CA [38]. Instead, we proposed a fully automatic method concatenating two NN models with no need for domain knowledge, or tuning, in the dataset considered.

We also performed an ablation study, eliminating the segmentation module of the proposed framework and finding the need to use the information provided by it to achieve accurate results. Additionally, a study was performed in order to gain insight into the effect on the overall performance of training a single CNN model or three independent networks to predict three target values. The results conclusively show that the original idea of training a single CNN with multiple outputs leads to better

results for atherosclerotic plaque detection.

Regarding the limitations of this study, highly unbalanced datasets with a small number of plaque images can be a problem in achieving results comparable to our proposal for the single CNN that uses only the original image as input. Additionally, in any medical application, explainability may be required. While the segmentation module of our proposal provides relevant information to the user, the classification and regression module is a black-box.

Our future research includes a study on other datasets to further confirm our conclusions, mainly in terms of generalization power. Therefore, we plan to revise this research work when more data from different institutions and different acquisition systems become available. This research paves the way for a fully automated evaluation of CIMT and plaque. We aim to make a more data from different institutions and different acquisition systems become available.

Furthermore, the method presented is a baseline framework to integrate information from the image and clinical data, which are very relevant to assess other pathologies or events related to atherosclerosis, such as cardiovascular risk. Moreover, since the output of the framework is for regression or classification targets, it can be easily adapted for medical purposes as cardiovascular risk prediction or risk stratification. In our future work, we intend to explore this field as another potential line of research.

Acknowledgements

This work was supported in part by the MICINN Grant RTI2018-095232-B-C21 and 2017 SGR 1742.

Conflict of interest statement

No potential competing interest was reported by the authors

Chapter 3

Deep Features for Risk Prediction

Deep-stratification of the cardiovascular risk by ultrasound carotid artery images

Maria del Mar Vila[1,2], **Lucas Gago**[3], **Pablo Pérez-Sánchez**[1,4], **Maria Grau**[5,6,7], **Beatriz Remeseiro**[8*], **Laura Igual**[3*]

[1]CIBER Enfermedades Cardiovasculares, Instituto de Salud Carlos III, Monforte de Lemos 3-5, Pabellón 11, 28029 Madrid, Spain

[2]IMIM, Institut Hospital del Mar d'Investigacions Mèdiques, Dr. Aiguader 88, 08003 Barcelona, Spain

[3]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain

[4]Institute for Biomedical Research of Salamanca (IBSAL), P. de San Vicente, 182, 37007 Salamanca, Spain

[5]Serra Húnter Fellow, Department of Medicine, Universitat de Barcelona, Carrer Casanova 143, 08036 Barcelona, Spain

[6]August Pi i Sunyer Biomedical Research Institute (IDIBAPS), C/ del Rosselló, 149, 08036 Barcelona, Spain

[7]CIBER Epidemiología y Salud Pública, Instituto de Salud Carlos III, Monforte de Lemos 3-5, Pabellón 11, 28029 Madrid, Spain

[8]Department of Computer Science, Universidad de Oviedo, Campus de Gijón s/n, 33203 Gijón, Spain

*These authors jointly supervised this work.

Abstract

Cardiovascular risk estimation functions predict the risk of cardiovascular events with clinical data and survival models. These functions accurately stratify individuals into low, moderate, and high-risk categories. However, they tend to classify a considerable number of individuals into the middle-risk category, and often, a subsequent reclassification into high-risk groups is required. Atherosclerosis is the leading cause of cardiovascular events, and ultrasound images of the Carotid Artery (CA) can detect its burden by measuring the carotid intima-media thickness and identifying atherosclerotic plaques. Current risk estimation functions do not consider ultrasound imaging. This paper proposes the use of *deep* ultrasound CA image features in survival models to improve risk stratification. In particular, we define new deep CA image features, extracting information from a convolutional neural network, and add them to an existing risk function. The experiments carried out show that using deep image features improves the AUC of the risk function to 0.842, and these features are enough to replace the information provided by blood biomarkers. Furthermore, the use of these features resulted in a 20% improvement in the reclassification of risk categories, specifically for individuals who suffered an event, as shown by the net reclassification improvement metric.

Keywords: Cardiovascular event, Survival Model, Reclassification, Convolutional neural networks, Machine learning

Resum

Les funcions d'estimació del risc cardiovascular prediuen el risc d'esdeveniments cardiovasculars amb dades clíniques i models de supervivència. Aquestes funcions estratifiquen amb precisió els individus en categories de risc baix, moderat i alt. No obstant això, tendeixen a classificar un nombre considerable d'individus en la categoria de risc mitjà i, sovint, es requereix una reclassificació posterior en grups d'alt risc. L'aterosclerosi és la causa principal dels esdeveniments cardiovasculars, i les imatges d'ecografia de l'artèria caròtida (AC) poden detectar-ne la càrrega mesurant el gruix íntima-mitjana carotidi i identificant plaques ateroscleròtiques. Les funcions actuals d'estimació del risc no tenen en compte la imatge per ecografia. Aquest article proposa l'ús de característiques profundes (*deep features*) de les imatges d'ecografia de l'AC en models de supervivència per millorar l'estratificació del risc. En particular, definim noves característiques profundes d'imatge de l'AC, extraient informació d'una xarxa neuronal convolucional,

i les afegim a una funció de risc existent. Els experiments realitzats mostren que l'ús de característiques profundes d'imatge millora l'AUC de la funció de risc fins a 0,842, i que aquestes característiques són suficients per substituir la informació proporcionada pels biomarcadors sanguinis. A més, l'ús d'aquestes característiques va donar lloc a una millora del 20% en la reclassificació de les categories de risc, específicament per als individus que van patir un esdeveniment, tal com mostra la mètrica de millora neta de la reclassificació (*net reclassification improvement*).

Paraules clau: Esdeveniment cardiovascular, Model de supervivència, Reclassificació, Xarxes neuronals convolucionals, Aprenentatge automàtic

3.1 Introduction

Cardiovascular Diseases (CVDs) are the leading cause of death in developed countries. Most at-risk individuals of cardiovascular events suffer from atherosclerosis, a chronic inflammatory process morphologically characterized by an asymmetric focal thickening of the innermost layer of the artery. Ultrasound Carotid Artery (CA) images are used to detect the burden of atherosclerosis since they provide the possibility to measure the Carotid Intima-Media Thickness (CIMT) (see Figure 3.1) of the artery and identify the presence of atherosclerotic plaques (Mannheim Consensus [8]). Monitoring the CIMT and detecting the atherosclerotic plaque, as well as its characteristics or changes, may have significant clinical relevance for CVD risk assessment.

In the field of cardiovascular epidemiology, practitioners use risk prediction functions [50, 51, 52, 53, 54] to estimate the risk of suffering an event in a period of time. These functions are based on survival models and estimate the risk using a set of clinical variables of each individual. Classically, the predictive results of risk functions are divided into different risk categories for stratification. Each risk category is defined by an interval probability of suffering a cardiovascular event in the next ten years. The cut-off points that define the risk stratification in *low*-risk group, *moderate*-risk group, and *high*-risk group have practical implications for deciding pharmacological intervention measures. Moreover, according to several clinical trials in primary prevention [55], treating populations is only cost-effective in the *high*-risk group.

One example of these risk prediction functions is REGICOR, which was built

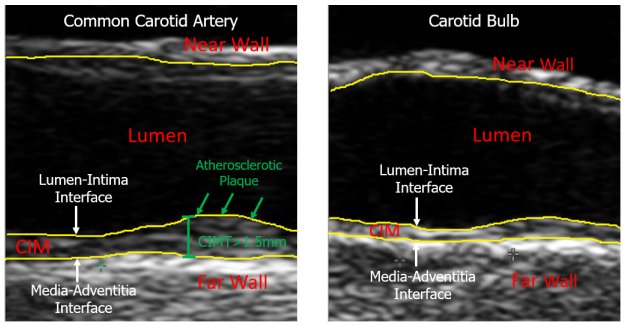

Figure 3.1: Ultrasound CA images from two territories: common CA (left) and bulb (right). The different parts of the CA are delimited with lines: Near wall, Far wall, Lumen, Bulb, Carotid Intima-Media (CIM) region. In both cases, the CIMT is estimated in the CIM region (between Lumen-Intima Interface and Media-Adventitia Interface). Atherosclerotic plaque is also shown (portion of the CIM with a CIMT greater than 1.5mm [8]).

based on REGICOR study [50] and has shown accurate predictions of cardiovascular events [50, 52, 53, 56]. REGICOR is a population-based observational cohort study conducted in the province of Girona (Spain). Table 3.1 shows data on the ten-year incidence rate of Ischemic Heart Disease (IHD) in a population of 3,724 individuals aged between 35 and 74 years who participated in REGICOR study. The risk category for each subject is based on the percentage probability of suffering a cardiovascular event in the next ten years (first column). Although the results shown in [50] are presented in four risk categories, we have summarized the table by combining the two middle categories (*low-moderate* and *high-moderate*) into a single category (*moderate*). The second and third columns indicate the number of subjects in each category and the number of events (IHD), respectively. The last column shows the percentage of events that occurred in a period of ten years for each risk category. As can be seen, the percentage of events falls inside each interval probability. More specifically, the percentage of subjects in the *low*, *moderate*, and *high* categories is 1.2%, 6.32%, and 12.50%, respectively, which is within the interval probability estimated: [0, 5)%, [5, 15)%, and [15,.)%. On the other hand, the values in the third column show that the *moderate* category concentrates the highest percentage of events (60%). This fact makes this stratification strategy ineffective in treating the population of the *moderate* group. Therefore, new pathological information should be considered to reclassify individuals from this group to the *high*-risk group.

In this research, we present a novel approach to improve the survival model for risk stratification proposed by Marrugat et al. [50]. In particular, our approach uses *deep*

Table 3.1: REGICOR data [50] on the ten-year incidence rate of Ischemic Heart Disease (IHD).

Risk category and interval probability	Subjects (%)	Subjects with IHD (%)	IHD in the risk group (%)
Low < 5	2449 (65.76%)	31 (25.83%)	1.27%
Moderate $\in [5, 15)$	1139 (30.59%)	72 (60.00%)	6.32%
High ≥ 15	136 (3.65%)	17(14.17%)	12.50%
TOTAL	3724 (100%)	120 (100%)	3.22%

ultrasound CA image features in the survival model aiming at reclassifying individuals from the *moderate*- to the *high*-risk category. In particular, we consider a deep neural network (DNN) architecture to extract a set of deep features from the CA images, add them to the REGICOR function, and analyze the new survival model in terms of prediction and reclassification. We show that the DNNs are able to learn new feature embeddings useful to improve risk stratification by classifying events from *low* or *moderate* categories to higher categories. To do so, we compare the performance of our model using different sets of deep features and with another model that uses a set of phenotypes manually defined from the CIM region. Additionally, we assess the relevance of these features by comparing the risk function outlined in [50] to a model with the same variables, except for the invasive ones (i.e., those requiring blood extraction), which are substituted with the image features. In this sense, we assess a model that does not include invasive variables but incorporates additional information about atherosclerosis, including details of the atherosclerotic plaque location. Finally, the best sets of image features are compared in terms of reclassification. In particular, we investigate whether the inclusion of deep image features in the survival model results in the reclassification of individuals who have experienced an event from the *moderate*-risk group to the *high*-risk group.

The rest of the manuscript is organized as follows. Section 3.2 exposes the related work. Section 3.3 presents the proposed methodology for cardiovascular risk estimation using deep image features. Section 3.4 describes the REGICOR dataset used in the experiments. Section 3.5 explains the experimentation setup and the performance measures used. Section 3.6 describes the experiments carried out and shows the corresponding results. Finally, Section 3.7 closes with the main conclusions and future lines of research.

3.2 Related Work

The tissue of CA walls provides information about the patients' arteries and cardiovascular health. For this reason, the study of ultrasound CA plaque images has been considered clinically relevant. Moreover, the long induction period of atherosclerosis makes it suitable for the study of subclinical CVD for preventive purposes. Several attempts in the literature tried to assess the cardiovascular risk of subjects using CA image features, as discussed below.

The main purpose of studies in [57, 58] was to create an image-based system to characterize the plaque and the carotid artery walls in ultrasound CA images. In these works, authors used the lumen diameter for risk stratification as ground truth to solve a classification problem between two ranges: low and high risk. Their proposal was to estimate the spatial distribution of gray levels to examine the texture for Common CA (CCA) far and near wall, reaching high classification accuracies with two classification methods: Support Vector Machine (SVM) and Principal Component Analysis (PCA).

Other works ([59, 60, 61, 62]) focused on analyzing the CA in longitudinal studies (i.e., repeated observations of the same subjects during a period of time). These works obtained CA image features, either manually or using semi-automatic methods, and used them to create risk prediction functions. Most of the image features considered are related to IMT (mean IMT, maximum IMT, minimum IMT, IMT variability) and atherosclerotic plaque (total plaque area, grayscale median of plaque). Since these features are used repeatedly in the literature, from now on, we will refer to this set of six features as *classical image phenotypes*. In particular, the model presented in [61] used the classical image phenotypes in two instants of time combined with conventional (non-imaging) risk variables achieving better results than classical survival models. Alternatively, Kyriacou et al. [63] used Probabilistic Neural Networks (PNN) and SVM classifiers to combine clinical features, phenotypes, and other CA image features (based on texture and morphology) for plaque classification (event vs non-event). Despite the promising results for predicting events in individuals with plaque, this classification task is different from the risk stratification problem we address in this study.

Moreover, several works in the literature [3, 64, 65, 4] compare classical risk prediction functions with different Machine Learning (ML) approaches. The main difference between both approaches is that ML methods use the event/non-event information for prediction (binary classification task) and survival models use the time until the event to predict the risk of suffering an event. In these studies, they used classical clinical variables, also known as *risk factors* and some additional characteristics

from a population initially free of cardiovascular disease in a longitudinal study. The ML techniques outperform existing approaches in cardiovascular event prediction. In this context, Weng et al. [3] compared different ML algorithms: random Forest (RF), logistic regression, gradient boosting, and neural networks, which reached the best results in cardiovascular prediction. The most recent approaches [64, 65] proposed boosted ensemble algorithms followed by automated feature selection using information gain ratio. Both of them reached very high accuracies in cardiovascular death [64] and cardiovascular event [65] prediction. Although they did not use CA images, Ambale-Venkatesh et al. [4] included features from electrocardiography images and reached the highest accuracy using the RF algorithm to predict the events and to select relevant features from a large set of variables (more than 700).

In terms of subject reclassification (i.e., moving subjects who suffered an event to higher risk categories and subjects free of event to lower risk categories), recent research works [66, 67] used Net Reclassification Improvement (NRI) to evaluate if the addition of one or more variables to a survival model improves its predictive capacity. In particular, these improvements are based on the reclassification of subjects with event into higher categories. Moreover, Tamarappoo et al. [65] applied this measure to evaluate the improvement in their proposed ML prediction model and reached an improvement of approximately 50% in reclassification.

Finally, it is worth mentioning some studies [68, 69],that characterize other types of information that affect cardiovascular health but are not conventional images or clinical variables. These studies deal with biomedical signals, such as PPG [68] or ECG [69]. The techniques in these two studies are, respectively, Hilbert transforms and wavelet transforms. The results achieved in both cases are very competitive in terms of signal characterization and detection of specific patterns. However, this is not the focus of our study.

3.3 Methodology

We propose to improve the survival model based on the REGICOR risk function [50] by combining its clinical variables with a novel set of deep features extracted from ultrasound CA images. Figure 3.2 depicts the different stages of the proposed method, which are subsequently described.

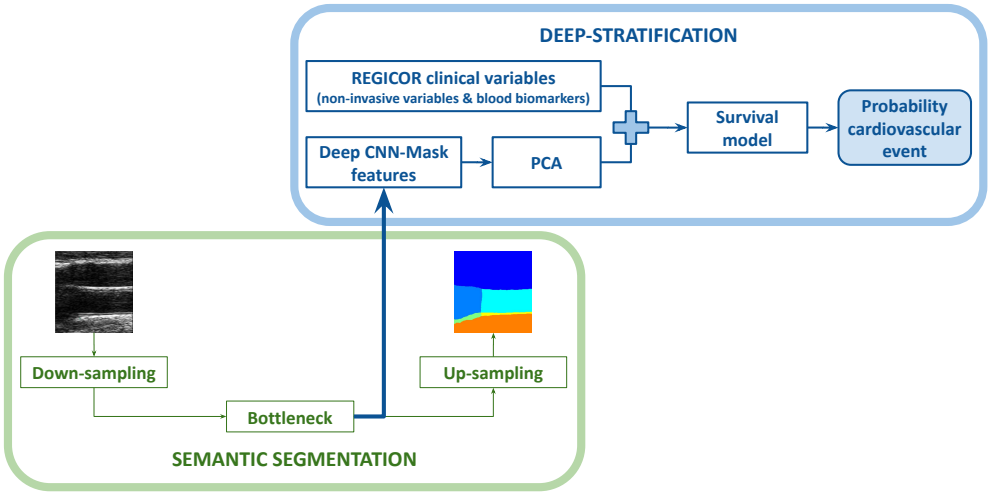

Figure 3.2: Proposed methodology for the deep-stratification of the cardiovascular risk. The survival model receives an input vector with 12 features, which include 8 clinical variables used in the REGICOR risk function and 4 deep CNN-Mask features extracted from a semantic segmentation model of the carotid intima-media [70] and transformed by PCA.

3.3.1 REGICOR Clinical Variables

Most of the risk prediction functions in the field of cardiovascular epidemiology are based on survival models, which are Cox Proportional Hazards models (CoxPh) [50, 51, 52, 54]. The CoxPh analyzes the risk affecting the survival of a population of subjects [71]. In these models, the outcome is the time until an event occurs, and the predictor covariates are the risk factors. In particular, CoxPh survival models, such as the ones used in Framingham [51] or in REGICOR [50], allow us to estimate the probability of suffering a cardiovascular event in the next period of time [71], say ten years, by means of the following formula:

$$prob(event|x) = 1 - S^{exp(\sum_{n=1}^{N} \beta_n x_n - \sum_{n=1}^{N} \beta_n \overline{x}_n)}, \tag{3.1}$$

where N is the number of risk factor variables, x_n is the value of the n-th variable for the individual x, β_n is the Cox coefficient for the n-th variable, $\overline{x}_n$ is the average value of the n-th variable in the validation population, and S is the average ten-year survival in the validation population. One of the best-known survival models in the literature is Framingham [51]. REGICOR risk function [50] is a model with the same risk factors as Framingham [51] but validated in the Spanish population. In particular, in the REGICOR risk function: $N = 8$, the average values correspond to the Spanish

population and S is the average ten-year survival in the Spanish population. The REGICOR risk function uses the following eight clinical variables:

- *Non-invasive variables*: age, sex, diastolic blood pressure, systolic blood pressure, and smoke.
- *Blood biomarkers*: diabetes, total cholesterol, and HDL cholesterol.

3.3.2 Deep CNN-Mask Features

The extraction of the deep features is based on the model defined by Gago et al. [70] for the segmentation of the carotid intima-media. Specifically, this work designed a semantic segmentation model that uses the U-Net [19], a widely recognized architecture that incorporates a downsampling path for feature extraction and an upsampling path for region segmentation. The downsampling path of the proposed model employs an EfficientNet B0 [42] as its backbone while maintaining the bottleneck and the upsampling path of the original U-Net.

The deep features used in our proposal are obtained from this semantic segmentation model. In particular, we use the deep features extracted from the bottleneck, which is located between the down-sampling and up-sampling paths (see Figure 3.2). As the segmentation model produces a mask of the CIM region, the 1,152 features derived in this manner are referred to as *deep CNN-Mask features*.

3.3.3 Dimensionality Reduction using Principal Component Analysis

The *one in ten rule* [72] used in statistics determines that the number of variables for a survival model should be around 1 variable for every 10 events. Taking into account the number of events in our study (3.22%, see Section 3.1) and the number of deep features (1,152) obtained for each territory (CCA and bulb), the dimension of the feature vector must be reduced.

For this purpose, we use Principal Component Analysis (PCA) [73], an algorithm that applies a linear transformation to compress a dataset onto a subspace of lower dimensionality while retaining the majority of the relevant information. The number of principal components to retain is typically determined by looking at the proportion of variance explained by each component and selecting a threshold, such as retaining the top n components that explain a certain percentage of the total variance.

In the current study, there are four CA images available for each subject: left and right sides of CCA and bulb territories. The deep CNN-Mask features are computed from the four images and then PCA is applied. The final vector per territory is calculated as the average of the feature vectors obtained from both sides (left and right). According to the experimentation, a variance percentage of 85% is applied to the two CCA images and 70% in the case of the bulb, thus obtaining two features per territory. That is, four deep features are concatenated with the eight REGICOR variables (which results in a final feature vector with $N = 12$), to feed the survival model, as following described.

3.3.4 Survival Model

At this point, we have obtained a 12-dimensional vector with the following features: eight clinical variables, which are the ones used in the REGICOR risk function; and four deep CNN-mask features, which are obtained after applying PCA to the deep features computed from a semantic segmentation model trained on ultrasound CA images from two territories (CCA and bulb). These 12 features feed the survival model, as shown in Equation 3.1, to finally estimate the probability of suffering a cardiovascular event in the next ten years. We call this strategy *deep-stratification* approach. In particular, for this work, we evaluate the respective CoxPh model (see Section 3.3.1), which estimates the time until the event. This model is obtained with `coxph()` function from Epi package [74] (R software [75]).

3.4 Dataset

This research work analyzes the participants from REGICOR [50], a population-based observational cohort study. The dataset consists of a sample of 5,083 subjects from *Girona's Heart Registry* [13] with ultrasound CA images and other clinical information. All these data were collected from 2007 to 2010 and the subjects represent a general population aged between 35 and 84 who were followed up for approximately ten years. Two trained sonographers performed the CA US scans with an Acuson XP128 US system equipped with L75-10 MHz transducer and a computer program extended frequency (Siemens-Acuson). US longitudinal images were obtained in B-mode with a resolution of 23.5 pixels/mm and a size of 470×445 pixels. The original images were saved in DICOM format and then converted to PNG. The set of images collected for each subject was obtained from left and right CA in two different territories (CCA and

bulb). During the CA acquisition, some images were discarded if the sonographers considered that the image quality was not sufficient. Due to the poor quality of the bulb images, they collected a total of 10,151 CCA images and 9,143 Bulb images. Among them, 4,727 CCA images and 3,721 bulb images have CIMT reference values, given by the Amsterdam Medical Center with a semi-automatic method [76].

The clinical data include eight classical risk factors used in the REGICOR risk function. We call them *REGICOR variables* and they are: gender, age, smoking, systolic and diastolic blood pressure, total cholesterol, HDL cholesterol, and diabetes. Moreover, these data include the time until the event or the time to the date of the last contact. Given the nature of our study, we only considered subjects with no history of CVD (coronary artery disease, stroke, or intermittent claudication), therefore 314 subjects were discarded. Finally, a total of 4,769 subjects were included in the dataset.

Table 3.2 shows a summary of the clinical data considered in this research. Note that, from the analyzed cohort of 4,769 subjects who were free from CVD at baseline, there are 151 incident cases (3.17%) with cardiovascular events (acute myocardial infarction, other ischemic heart diseases, stroke, and the same causes of death) during the ten-year follow-up period.

Table 3.2: Summary of clinical data of the REGICOR subjects considered in this study, grouped by 'sex' and with the p-value for the differences between the two groups. Categorical variables are expressed as *n (%)* and continuous variables as *mean (standard deviation)*.

	All N=4,769	Men N=2,620	Women N=2,149	p.value
Age	59.5 (11.8)	59.7 (11.8)	59.4 (11.8)	0.483
Total cholesterol	205 (35.5)	209 (35.2)	201 (35.3)	<0.001
HDL cholesterol	52.7 (11.8)	56.6 (11.6)	48.0 (10.3)	<0.001
Systolic blood pressure	129 (19.7)	126 (20.1)	133 (18.2)	<0.001
Diastolic blood pressure	77.2 (10.1)	75.1 (9.77)	79.7 (9.91)	<0.001
Diabetes	608 (12.7%)	252 (9.62%)	356 (16.6%)	<0.001
Smoke	810 (17.1%)	347 (13.3%)	463 (21.7%)	<0.001
Event	151 (3.17%)	58 (2.21%)	93 (4.33%)	<0.001

3.5 Experimental Setup

This section describes the setup used in the experimentation carried out to evaluate our deep-stratification approach. First, we explain the performance measures to evaluate the prediction capability of the survival model and the reclassification. Next, we

describe the train-test split applied to the REGICOR dataset for validation purposes.

Note that the code of our proposed framework will be publicly available after paper acceptance.[1]

3.5.1 Evaluation Metrics

In order to evaluate the performance of our deep-stratification approach we used two different metrics: the Area Under the Curve (AUC) and the Net Reclassification Improvement (NRI).

The AUC metric is used to assess the performance of the survival model. In this context, this metric agrees with Harrell's C [77], which is a goodness of fit measure of risk scores models such as statistical survival models.

On the other hand, NRI [78] is used to evaluate the reclassification improvement in risk prediction. NRI is a widely used metric in the recent clinical literature [66, 67, 65], which intuitively summarizes the improvement in the classification of individuals in risk categories. Therefore, it is used to summarize the incremental improvement obtained with new variables. The total improvement is the sum of the improvements in the reclassification of individuals without event and with event separately as is shown in the following formula:

$$NRI = (N1 - N2) + (N3 - N4) \tag{3.2}$$

where $N1$ and $N2$ are the percentages of individuals who have suffered an event and those who have been reclassified into higher and lower categories, respectively; and $N3$ and $N4$ are the percentages of individuals free of event who have been reclassified into lower and higher categories, respectively.

Note that when the problem is the probability of suffering an event during a period of time, we use the event variable (*yes/no*) and the time until the event (or the time until the date of the last contact in case of non-event subjects). For this reason, the statistical techniques utilized for NRI, such as the `nricens` function from `nricens`[79] package (R software [75]), which is employed in this study, incorporate survival functions that consider the time to the event in addition to the values shown in the formula. That is, these functions generate an NRI output that cannot be derived directly from the aforementioned formula 3.2.

[1] `https://github.com/mmarvila/CA_deep_stratification`

3.5.2 Train-test Split

Table 3.3 shows the number of images used to evaluate our deep-stratification approach. As explained in Section 3.4, the REGICOR dataset contains a total of 10,151 CCA images and 9,143 bulb images. Among them, 4,727 CCA images and 3,721 bulb images have CIMT reference values that can be used as the Ground Truth (GT) for CIMT estimation. Note that the first two rows of the table (CIMT GT) correspond to the same train-validation-test split used in [70] (60% training set, 20% validation set, and 20% test set), which has been used here to evaluate image-based features.

Table 3.3: Number of images in the REGICOR dataset: train-validation-test split to evaluate the image-based features (CIMT GT) and test split to evaluate the survival models (CIMT GT + NO CIMT GT).

		TRAIN	VALIDATION	TEST	TOTAL
CIMT GT	CCA	2836	946	945	4727
	BULB	2232	744	745	3721
NO CIMT GT	CCA			5424	5424
	BULB			5422	5422
TOTAL	CCA	2836	946	6369	10151
	BULB	2232	744	6167	9143

The images without CIMT values can be used for the evaluation of the survival model. For this reason, we have also considered them and, therefore, the test set used in this case is composed of 6,369 CCA and 6,167 bulb images.

3.6 Results

This section details the four experiments conducted to evaluate our deep-stratification approach and analyzes the results obtained. The first two experiments are intended to select the best set of image-based features. In particular, we compare two sets of deep features in Experiment 1 and a set of hand-crafted features in Experiment 2. In both cases, we consider two territories (CCA and bulb) and several sizes of feature vectors. Next, in Experiment 3, we assess the relevance of the image-based features with respect to the blood biomarkers. For this purpose, we compare the survival model of the REGICOR function [50] but with the image-based features instead of the blood biomarkers. Finally, in Experiment 4, we analyze the most competitive configurations of the previous experiments in terms of the reclassification of the survival model.

The computational time and complexity of the proposed methodology mostly depends on the model used to extract deep features. As mentioned in [37], the time needed to process an input image is 0.026 seconds using a GeForce RTX 2080ti 11GB GPU also from NVIDIA. On the other hand, the computational time to obtain the proposed CoxPh survival model is less than one millisecond per individual.

3.6.1 Experiment 1: Analysis of the Deep Features

In order to evaluate the adequacy of the proposed deep CNN-Mask features, we compare them with another set of deep features. For the extraction of the alternative deep features, we use a CNN model defined for CIMT estimation and plaque detection in [70]. This model was tuned using Bayesian optimization, with a final architecture composed of four blocks of convolutional layers followed by a max-pooling operation and two blocks of fully connected layers followed by batch normalization and dropout. In this case, the deep features were extracted from the second block of fully connected layers. The 32 features derived in this manner are referred to as *deep CNN-CIMT features*.

As we mentioned in Section 3.4 the CA images of the REGICOR dataset correspond to two territories, CCA and bulb. The quality of the images in CCA is better, whilst there is an increased burden of atherosclerotic plaque in the bulb images. For this reason, we compute the deep features from both territories, individually and jointly for comparison.

For each CA image, the size of the feature vector is 1,152 for the proposed deep CNN-Mask features and 32 for the alternative deep CNN-CIMT features. Since the number of events in the dataset is small (151 from 4,769 subjects, see Table 3.2), the size of the feature vectors must be reduced. As previously explained in Section 3.3.3, the number of events in the dataset should be about 10 per risk factor variable in the survival model [72]. Thus, we set the maximum number of variables to be included in the survival model to 16. Taking into account that our proposed model already includes the 8 REGICOR risk function variables [50] (see Figure 3.2), the number of deep features should be less than or equal to 8. To reduce the size of the two feature vectors considered, we apply PCA to obtain a maximum of 8 variables and keep at least 70% of the variance. Table 3.4 shows the number of features obtained after applying PCA with different variance percentages. For CCA images, a variance of 99% in deep CNN-Mask features and 90% in CNN-CIMT features result in feature vectors with more than 8 variables, so lower variances must be considered for the proposed

survival model. Regarding the other territory, bulb, a variance lower than 95% must be considered in both feature sets.

Table 3.4: Kept variance and number of features obtained when applying PCA to the two sets of deep features.

	CNN-Mask		CNN-CIMT	
Territory	Var.	No. feats.	Var.	No. feats.
CCA	99%	>8		
	95%	4	90%	>8
	85%	2	85%	8
	70%	1	70%	5
BULB	95%	>8	95%	>8
	90%	5	90%	6
	85%	4	85%	4
	70%	2	70%	2

As detailed in Section 3.3, each subject has two CA images (left and right) per territory (CCA and bulb). Thus, the final feature vector is calculated as the average of the feature vectors extracted from both sides. In case one of the two images is missing, the subject is discarded. For this reason, using the 6,369 CCA and 6,167 bulb images mentioned in section 3.5.2, we obtain the deep features for the following number of subjects: 2,796 subjects for CCA, 2,760 subjects for bulb, and 2,501 for both territories.

Table 3.5 shows the performance of the survival model using the two sets of deep features (CNN-Mask and CNN-CIMT), in the two territories (CCA and bulb) individually and jointly, and with different kept variances for PCA, based on the results reported in Table 3.4. In all cases, the deep features are concatenated with the 8 REGICOR variables [50] that are used as the baseline (AUC = 0.825). For each configuration, we report the total number of features (No. feats.), the predictive capacity of the survival model (AUC), the AUC increase with respect to the baseline, and the p-value of the AUC increase, which is considered statistically significant if $p < 0.06$. Note that equal values in the AUC metric have different increments due to rounding precision (three decimal digits), and vice versa.

As can be observed, the AUC increase reaches statistically significant values using either of the two sets of deep features. However, the best AUC results are obtained with the deep CNN-Mask features, with AUC greater than 0.840 in several configurations. With respect to the two territories, their combination provides the highest values when the deep CNN-Mask features are applied. Regarding the number of features, the highest AUC values are obtained with larger feature vectors, mainly because

Table 3.5: Experiment 1. AUC results of the survival model fed with the 8 REGICOR variables and different sets of deep features (CNN-Mask and CNN-CIMT), applied to the input images of two territories (CCA and bulb). The number of features obtained after applying PCA to the deep features is specified between parentheses (F), after the variance percentage. The statistically significant results ($p < 0.06$) are in bold.

Deep features	Territory		No. feats.	AUC	AUC increase †	p
	CCA	Bulb				
CNN-Mask	PCA 95% (4F)	-	12	0.831	0.007	0.118
CNN-Mask	PCA 85% (2F)	-	10	0.831	0.006	0.230
CNN-Mask	PCA 70% (1F)	-	9	0.827	0.002	0.628
CNN-Mask	-	**PCA 90% (5F)**	**13**	**0.842**	**0.018**	**0.050**
CNN-Mask	-	**PCA 85% (4F)**	**12**	**0.842**	**0.017**	**0.052**
CNN-Mask	-	PCA 70% (2F)	10	0.834	0.010	0.296
CNN-Mask	**PCA 95% (4F)**	**PCA 85% (4F)**	**16**	**0.847**	**0.023**	**0.008**
CNN-Mask	**PCA 85% (2F)**	**PCA 90% (5F)**	**15**	**0.846**	**0.022**	**0.016**
CNN-Mask ⋆	**PCA 85% (2F)**	**PCA 70% (2F)**	**12**	**0.842**	**0.017**	**0.054**
CNN-CIMT	PCA 85% (8F)	-	16	0.833	0.008	0.076
CNN-CIMT	**PCA 70% (5F)**	**-**	**13**	**0.832**	**0.008**	**0.046**
CNN-CIMT	-	PCA 90% (6F)	14	0.835	0.011	0.198
CNN-CIMT	-	PCA 85% (4F)	12	0.836	0.011	0.070
CNN-CIMT	-	PCA 70% (2F)	10	0.834	0.010	0.296
CNN-CIMT	PCA 70% (5F)	PCA 70% (2F)	15	0.836	0.011	0.066

⋆ Our proposed method.
† AUC increase with respect to the 8 REGICOR variables [50] (baseline).

these configurations correspond to the joint use of the two territories. Note that the configurations that achieve statistically significant results ($p < 0.06$) are selected for further analysis (Experiments 3 and 4).

3.6.2 Experiment 2: Analysis of the Hand-crafted Features

The aim of this experiment is to analyze the use of a new set of *hand-crafted features*, which could replace the deep features in our proposed methodology (see Figure 3.2). Based on the classical image phenotypes mentioned in Section 3.2, we consider six phenotypes manually defined from the CIM region (see Section 3.1). Thus, for each CA image we obtain the following image features: mean IMT, maximum IMT, minimum IMT, IMT variability, Total Plaque Area (TPA), and Grayscale Median (GSM) of plaque. Notice that TPA is measured in mm^2 and it is estimated in the region where the CIMT reaches more than 1.5 mm, following the Mannheim consensus [8]. GSM refers to the grayscale median value in the same area where TPA is evaluated.

These phenotypes are extracted from the four CA images available for each subject (left and right sides of CCA and bulb territories), thus obtaining a total of 24 phenotypes per subject. As previously explained in Section 3.6.1, a maximum of 8 image-based features should be added so a dimensionality reduction procedure must be applied. The definition of these 24 hand-crafted features is based on prior knowledge about

specific characteristics of the carotid images. Therefore, we can reduce the number of features, from 24 to a maximum of 8, by selecting the most relevant phenotypes according to the results of a statistical analysis performed on the training data. The procedure carried out is summarized as follows:

- **Base-e logarithmic adjustments.** Some phenotypes are not normally distributed and hence they should be normalized using base-e logarithmic adjustment. These phenotypes are mean IMT, maximum IMT, and IMT variability (on both sides).

- **Categorization of variables.** Due to the low percentage of plaques (3.4% in CCA and 25.8% in bulb), TPA and GSM phenotypes are categorized into three classes. The categories for TPA are *non-plaque*, *small plaque*, and *high burden of plaque*, where the threshold between the last two categories is the median of all TPA values from the same side (five for CCA and six for bulb). GSM is categorized into *non-plaque*, *echolucent*, and *non-echolucent*, where the threshold for echolucency detection is the third quartile of all GSM values from the same side (107 in CCA and 95 in bulb). Finally, in order to analyze the interaction between the TPA and the GSM phenotypes, we create a new variable as a combination of both. The categories of this new variable are *non-plaque* (the subject does not have any plaque), *small plaque and non-echoic* (the subject has at least plaque in one side of the arteries, but none is huge or echoic), *small plaque and echoic* (the subject has at least plaque in one side that is echoic, but there is not any huge plaque), *huge plaque and non-echoic* (the subject has at least a huge plaque that is not echoic), and *huge plaque and echoic* (the subject has at least a huge plaque that is echoic). This categorization is done for CCA phenotypes and bulb phenotypes separately. Finally, since there is no event in the *small plaque and non-echoic* category for CCA, we merge it with the *non-plaque* category.

- **Mean between left and right sides.** In order to reduce the number of phenotypes, the mean IMT, maximum IMT, minimum IMT, and IMT variability are defined as the average of right and left side values [13]. If a value is missing on one side, we use the available value for this subject. If the value is missing on both sides, then the phenotype is considered missing.

- **Co-linearity.** We eliminate co-linear variables using the Variance Inflation Factor (VIF)[80]. In particular, we analyze all the variables from the model and

discard the ones with VIF>2. Maximum IMT phenotypes from CCA and bulb are also discarded for the model.

- **Discarding variables.** Ultimately, our approach involves systematically eliminating non-statistically significant phenotypes from the survival model. However, we make a deliberate choice to retain the phenotype variables that are deemed confounders. Note that a variable is considered a confounder if at least one of the coefficients of the variables that remained in the model changed more than 15%. After the analysis, all the selected phenotypes are considered confounders so we keep all of them.

As a result, Table 3.6, shows the coefficients for the CoxPh model (see Section 3.3.1) and their corresponding p-value. These coefficients correspond to a survival model including the eight REGICOR variables and the eight phenotypes selected. Note that the variable is statistically significant for the model if $p < 0.06$.

Table 3.6: Coefficients for the CoxPh model and p-values of the risk factors used in the survival model: eight factors from the REGICOR risk function and the six hand-crafted phenotypes selected based on the statistical analysis performed.

Risk factor	Territory	Coefficient	p
Age	-	0.06	<0.01
Sex	-	0.38	0.06
Total cholesterol	-	0.01	0.04
HDL cholesterol	-	-0.03	0.00
Systolic blood pressure	-	0.00	0.62
Diastolic blood pressure	-	0.01	0.23
Diabetes	-	0.49	0.02
Smoker	-	0.12	0.66
log(mean IMT)	CCA	2.55	<0.01
minimum IMT	CCA	-1.26	0.26
log(IMT Variability)	CCA	-0.43	0.17
Small plaque and echoic	CCA	-0.61	0.32
Huge plaque and non-echoic	CCA	0.36	0.59
Huge plaque and echoic	CCA	-0.70	0.25
log(mean IMT)	Bulb	-1.48	0.30
minimum IMT	Bulb	0.22	0.83
log(IMT Variability)	Bulb	0.71	0.14
Small plaque and echoic	Bulb	0.15	0.61
Huge plaque and non-echoic	Bulb	0.13	0.78
Huge plaque and echoic	Bulb	0.26	0.44

Table 3.7 shows the performance of the survival model using the hand-crafted

features in the two territories (CCA and bulb), individually for the Statistical Analysis (SA) above described and jointly for PCA (maximum 8 variables and at least 70% of variance). In all cases, the hand-crafted features were concatenated with the 8 REGICOR variables [50] that are used as the baseline (AUC = 0.825). As in Experiment 1, we report the total number of features (No. feats.), the predictive capacity of the survival model (AUC), the AUC increase with respect to the baseline, and the p-value of the AUC increase, which is considered statistically significant if $p < 0.06$. Note that equal values in the AUC metric have different increments due to rounding precision (three decimal digits), and vice versa.

Table 3.7: Experiment 2. AUC results of the survival model fed with the eight REGICOR variables and the hand-crafted features applied to the input images of two territories (CCA and bulb). SA stands for statistical analysis and PCA is followed by the variance percentage applied. In both cases, the number of features obtained is specified between parentheses (F). The statistically significant results ($p < 0.06$) are in bold.

Image features	Territory		No. feats.	AUC	AUC increase †	p
	CCA	Bulb				
Hand-crafted	**SA (4F)**	-	12	**0.843**	0.018	**0.010**
Hand-crafted	-	SA (4F)	12	0.839	0.015	0.116
Hand-crafted	**SA (4F)**	**SA (4F)**	16	**0.856**	0.031	**0.004**
Hand-crafted	PCA 99% (5F)		13	0.830	0.006	0.278
Hand-crafted	PCA 95% (4F)		12	0.826	0.001	0.734
Hand-crafted	PCA 90% (3F)		11	0.824	-0.001	1.314
Hand-crafted	PCA 80% (2F)		10	0.824	<0.000	1.280

† AUC increase with respect to the 8 REGICOR variables [50] (baseline).

As can be seen, the AUC increase is statistically significant when using the features selected by the statistical analysis on the CCA territory, both individually or combined with the bulb. In particular, the best performance is achieved when using the two territories jointly. On the contrary, the configurations that used the feature vectors obtained after applying PCA to the hand-crafted features do not have statistical significance. Note that the configurations that achieve statistically significant results ($p < 0.06$) are selected for further analysis (Experiments 3 and 4).

3.6.3 Experiment 3: Analysis of the REGICOR Variables

The aim of this experiment is to analyze the power of image-based features and see if it is possible for them to replace the 3 blood biomarkers used in the REGICOR risk function [50].

Table 3.8 shows the performance of the survival model using the 5 non-invasive REGICOR variables concatenated with the different configurations of image-based

features selected in the previous experiments, according to their statistical significance. Note that the target of the experiment is to analyze if these features can replace the 3 blood biomarkers, so the 8 REGICOR variables are used as the baseline (AUC = 0.825). As in previous experiments, we report the total number of features (No. feats.), the predictive capacity of the survival model (AUC), the AUC increase with respect to the baseline, and the p-value of the AUC increase, which is considered statistically significant if $p < 0.06$. Note that equal values in the AUC metric have different increments due to rounding precision (three decimal digits), and vice versa.

Table 3.8: Experiment 3. AUC results of the survival model fed with the five non-invasive REGICOR variables and different configurations of image-based features selected in Experiments 1 and 2, applied to the input images of two territories (CCA and bulb). SA stands for statistical analysis and PCA is followed by the variance percentage applied. In both cases, the number of features obtained is specified between parentheses (F).

Image features	Territory		No. feats.	AUC	AUC increase †	p
	CCA	Bulb				
CNN-Mask	-	PCA 90% (5F)	10	0.826	0.002	0.894
CNN-Mask	-	PCA 85% (4F)	9	0.826	0.002	0.892
CNN-Mask	PCA 95% (4F)	PCA 85% (4F)	13	0.831	0.007	0.57
CNN-Mask	PCA 85% (2F)	PCA 90% (5F)	12	0.831	0.006	0.63
CNN-Mask ⋆	PCA 85% (2F)	PCA 70% (2F)	9	0.827	0.002	0.874
CNN-CIMT	PCA 70% (5F)	-	10	0.813	-0.012	1.716
Hand-crafted	SA (4F)	-	9	0.825	0.000	0.978
Hand-crafted	SA (4F)	SA (4F)	13	0.839	0.014	0.352

⋆ Our proposed method
† AUC increase with respect to the 8 REGICOR variables [50] (baseline).

As can be observed in Table 3.8, there is a positive increment in most of the survival models fed with image-based features with respect to the survival model fed with the three blood biomarkers (REGICOR risk function [50]). Note that only in the case of the deep CNN-CIMT features the AUC increment is negative.

3.6.4 Experiment 4: Analysis of the Reclassification Results

This experiment aims at analyzing the models selected in Experiments 1 and 2 in terms of their reclassification results for cardiovascular events. For this purpose, we consider the NRI metric (see Section 3.5.1) and the following cut-off points, which correspond to the risk categories defined in [50] and discussed in Table 3.1:

- *low*: Subjects with a probability of suffering an event <0.05.
- *low-moderate*: Subjects with a probability of suffering an event in the range $[0.05, 0.1)$.

- *high-moderate*: Subjects with a probability of suffering an event in the range $[0.1, 0.15)$.
- *high*: Subjects with a probability of suffering an event ≥ 0.15.

Table 3.9 shows the performance of the survival model using the 8 REGICOR variables [50] concatenated with the different configurations of image-based features selected in Experiments 1 and 2, according to their statistical significance. For each configuration, we report the total NRI value (NRI) and its Confidence Interval (CI), the NRI value for the subjects who suffered an event (NRI events) and its CI, and the NRI value for the subjects free of event (NRI controls) and its CI. Note that the NRI values are statistically significant if their CI does not include 0 and they are shown in bold.

Table 3.9: Experiment 4. NRI results of the survival model fed with the 8 *REGICOR variables* and the different configurations of image-based features selected in Experiments 1 and 2, applied to the input images of two territories (CCA and bulb). SA stands for statistical analysis and PCA is followed by the variance percentage applied. In both cases, the number of features obtained is specified between parentheses (F). The statistically significant results (CI does not include 0) are in bold.

Image features	Territory		NRI [CI 95%]	NRI events [CI 95%]	NRI controls [CI 95%]
	CCA	Bulb			
CNN-Mask	-	PCA 90% (5F)	11.54 [-0.05;0.30]	9.99 [-0.07;0.28]	1.55 [0.00;0.03]
CNN-Mask	-	PCA 85% (4F)	11.56 [-0.05;0.30]	10.05 [-0.07;0.28]	1.52 [0.00;0.03]
CNN-Mask	PCA 95% (4F)	PCA 85% (4F)	**17.91 [0.01;0.34]**	16.1 [-0.01;0.32]	**1.81 [0.01;0.03]**
CNN-Mask	PCA 85% (2F)	PCA 90% (5F)	16 [-0.01;0.32]	14.76 [-0.02;0.31]	1.24 [<0.00;0.03]
CNN-Mask ⋆	PCA 85% (2F)	PCA 70% (2F)	**20.82 [0.04;0.38]**	**20.02 [0.03;0.37]**	0.81 [-0.01;0.02]
CNN-CIMT	PCA 70% (5F)	-	**17.50 [0.03;0.33]**	**17.14 [0.03;0.32]**	0.36 [-0.01;0.01]
Hand-crafted	SA (4F)	-	9.79 [-0.03;0.21]	8.85 [-0.04;0.20]	0.95 [<0.00;0.02]
Hand-crafted	SA (4F)	SA (4F)	6.47 [-0.11;0.26]	5.58 [-0.12;0.25]	0.9 [-0.01;0.02]

⋆ Our proposed method

The results presented in Table 3.9 indicate statistically significant improvements in reclassification for three specific configurations. These improvements were achieved using deep features, resulting in an increase of more than 17%. Only these three configurations show statistically significant results in either the "NRI events" or in the "NRI controls" column. Particularly, with our proposal, we show a significant increment in subjects who suffered an even, with an NRI of 20.02%, while the increment using CNN-CIMT features is lower, 17.14%. Instead, the other configuration using

CNN-Mask shows a statistically significant increment in subjects free of event, although it is relatively small (1.8%). In contrast, the results obtained from the hand-crafted features do not demonstrate statistical significance in any case.

Table 3.10 shows the reclassification of the event group using the previously defined risk categories. The risk categories and the number of subjects are shown using the REGICOR risk function [50] (rows) and our proposal (columns). The values above the diagonal indicate subjects that have been assigned a higher category compared to REGICOR [50]. Conversely, values below the diagonal represent the number of subjects that have been downgraded. Similarly, the values on the diagonal indicate the number of subjects classified with the same category using both our proposal and REGICOR [52]. Although here we cannot show the calculation of the NRI values (as we mentioned in Section 3.5.1) Table 3.10 shows that, for the events, there are many more individuals who are reclassified into higher categories (17=2+5+6+4) than not than lower categories (4=1+1+1+1).

Table 3.10: Comparison between the REGICOR risk function [50] and our proposed method in terms of reclassification results for cardiovascular events.

REGICOR	Our proposed method				
risk function	<0.05	$[0.05, 0.1)$	$[0.1, 0.15)$	≥ 0.15	TOTAL
<0.05	24	2	0	0	26
$[0.05; 0.1)$	1	4	5	6	16
$[0.1; 0.15)$	1	1	2	4	8
≥ 0.15	0	0	1	9	10
TOTAL	26	7	8	19	60

3.6.5 Comparison with the Literature

The predictive capacity of our proposed method (AUC = 0.842, Table 3.5) is similar to or higher than the results proposed in the literature (see Section 3.2). Even without the 3 blood biomarkers used in the REGICOR risk function, we achieved a good cardiovascular risk prediction (AUC = 0.827, Table 3.8).

The best result found in the literature is reported in [61] (AUC = 0.93). This study uses risk factors and phenotypes at two different times, thus making data collection more complex. Regarding the ML approaches, we do not reach the results reported in [4] (AUC = 0.86), but their proposed survival model includes data from questionnaires (lifestyles, history, medication, etc.), more biomarkers, electrocardiography, and magnetic resonance imaging features. That is, they use a set of characteristics that

make the study more expensive and that are not appropriate to be obtained in primary care centers. Regarding deep features, our best result is with "PCA 95%, CCA & 85%, BULB" features, which reach 0.847 in the AUC metric (Table 3.5. In the literature, we also found the study conducted by Rine Nakanishi [64] who demonstrates a slightly superior AUC result of 0.85. However, they do use computed tomography images, a more expensive technique than US imaging.

In terms of reclassification, the total NRI reported by Tamaroppoo et al. [65] (53%) is better than the one obtained with our method (20.82%, Table 3.9). However, our method outperforms it in the event group, which is the objective of the presented proposal due to its clinical relevance. More specifically, the NRI of the event group reported in [65] is 8% versus 20.02% achieved with our proposal.

3.7 Conclusions

This work presents, for the first time in the literature, a survival model that integrates CA image features extracted from deep neural networks. The new survival model is capable of predicting the risk of suffering a cardiovascular event, which results in a deep-stratification of the cardiovascular risk. The proposal improves the survival model presented in [50] by adding information from CA ultrasound images. For that, we concatenated deep features from a CNN previously defined for CA image semantic segmentation. These features are able to improve the model in terms of prediction (AUC=0.84, with an increment of 0.017 with respect to REGICOR risk function [50]) and reclassification (NRI=20.8%, NRI events=20%). We successfully achieved a reduction in the number of individuals in the middle-risk category and moved them to the *high*-risk category, which is our main goal.

In order to validate our proposal, we performed a comparison of different sets of image features and different configurations of these features. First, we compared our proposal with another set of deep features that reached a statistically significant improvement in prediction and reclassification, but with a smaller increase than with our proposal (AUC=0.83, NRI= 17.5%, and NRI events= 17.1%). Second, we compared with a set of phenotypes manually defined from the CIM region and selected according to the statistical analysis results. In this case, the improvement reached in prediction is high (AUC=0.86) and statistically significant, but the findings in reclassification were not statistically conclusive. In addition, our findings demonstrate that CA image features are able to replace invasive variables, such as blood biomarkers, while simultaneously providing localized information concerning atherosclerotic plaque.

The main limitation of our work is the small number of events in the dataset (151 events over 4769). With so few events, we are forced to greatly reduce the number of new image features and we are also exposed to overfitting. In this scenario, the survival models may not have enough statistical power to show all the differences between the different sets of image features. In addition, the proposed method has been tested on a single dataset, so it has not been possible to analyze its generalizability.

For future research, it would be useful to validate our method with an independent cohort that has more events. This would allow us to overcome the aforementioned limitations, increasing the statistical significance of the study and testing the power of generalization of our approach. In addition, it would be interesting to use other deep features obtained with CNNs combined with other deep neural networks trained for another task, such as a binary classification task (to predict event/non-event) or a regression task (to estimate the time until the event). Another line of research could be the interpretability of the specific features extracted from CA images. Understanding the contribution of individual features to the survival model, for example generating saliency maps, can provide information about which regions in CA images are associated with cardiovascular risk. Finally, it would be interesting to perform a longitudinal analysis of the deep features, including changes over time in CA, as it is suggested in [61], which uses risk factors and phenotypes at two different times.

Acknowledgments

This work was partially supported by the MICINN Grant RTI2018-095232-B-C21, AGAUR Grants (2021-SGR-01104, 2017-SGR-222 and 2017-SGR-1742), and the Spanish Ministry of Economy and Competitiveness through the Instituto de Salud Carlos III-FEDER (CIBERCV and FIS CPII17/00012).

Chapter 4

3D Cardiovascular Segmentation

Context-Aware Multilevel EfficientNet-UNet++ for Precise 3-D Carotid Vessel-Wall Segmentation

Lucas Gago[1], **Beatriz Remeseiro**[2*], **Laura Igual**[1*]

[1]Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona 08007, Spain

[2]Department of Computer Science, Universidad de Oviedo, Campus de Gijón s/n, Gijón 33203, Spain

*These authors jointly supervised this work.

Abstract

Accurate measurement of the carotid artery vessel wall thickness is paramount for monitoring patients with atherosclerosis. In 3D Magnetic Resonance (MR) imaging, achieving precise vessel wall segmentation is essential, but segmenting the entire image is both challenging and time-consuming. This paper presents an innovative deep learning segmentation approach that offers a practical and accurate solution for measuring carotid artery vessel wall thickness in MR images, thus potentially improving the monitoring and treatment of patients with atherosclerosis. We propose a multilevel EfficientNet-UNet++ network operating on a slice-by-slice basis and including contextual information. The approach entails four fundamental steps at

different levels: (1) extracting the non-black region from the MR image, (2) identifying the region of interest (ROI) using the local 3D context, (3) upscaling the area around the segmented ROI, focusing specifically on the vessel wall and surrounding tissue, and (4) re-segmenting the upscaled ROI to extract finer details of the vessel. Our multilevel segmentation approach outperforms state-of-the-art results in a public dataset comprising 2718 annotated slices from 75 patients. The experimentation and analysis demonstrate its effectiveness and potential value in clinical applications, with precise results near arterial bifurcation and in the presence of atherosclerotic plaque. The proposed fully automated method, based on the combination of two powerful neural network architectures, provides a reliable and accurate solution to segment the carotid artery vessel wall in MR images.

Keywords: Carotid outline, Semantic Segmentation, Atherosclerotic Plaque, Deep Learning, EfficientNet, UNet++

Resum

La mesura precisa del gruix de la paret del vas de l'artèria caròtida és fonamental per al monitoratge de pacients amb aterosclerosi. En la imatge per Ressonància Magnètica (RM) 3D, aconseguir una segmentació precisa de la paret del vas és essencial, però segmentar tota la imatge és un repte i requereix molt de temps. Aquest article presenta un enfocament innovador de segmentació basat en aprenentatge profund (deep learning) que ofereix una solució pràctica i precisa per mesurar el gruix de la paret del vas de l'artèria caròtida en imatges de RM, millorant així potencialment el monitoratge i tractament dels pacients amb aterosclerosi. Proposem una xarxa EfficientNet-UNet++ multinivell que opera tall per tall (slice-by-slice) i inclou informació contextual. L'enfocament comporta quatre passos fonamentals a diferents nivells: (1) extreure la regió no negra de la imatge de RM, (2) identificar la regió d'interès (ROI) utilitzant el context 3D local, (3) escalar l'àrea al voltant de la ROI segmentada, centrant-se específicament en la paret del vas i el teixit circumdant, i (4) resegmentar la ROI escalada per extreure detalls més fins del vas. El nostre enfocament de segmentació multinivell supera els resultats de l'estat de l'art en un conjunt de dades públic que comprèn 2.718 talls anotats de 75 pacients. L'experimentació i l'anàlisi demostren la seva eficàcia i valor potencial en aplicacions clíniques, amb resultats precisos prop de la bifurcació arterial i en presència de placa ateroscleròtica. El mètode proposat, totalment automàtic i es basa en la combinació de dues potents arquitectures de xarxes

neuronals, proporciona una solució fiable i precisa per segmentar la paret del vas de l'artèria caròtida en imatges de RM.

Paraules clau: Contorn carotidi, Segmentació semàntica, Placa ateroscleròtica, Aprenentatge profund, EfficientNet, UNet++

4.1 Introduction

Atherosclerosis, once perceived as a major health problem primarily in Western countries, has now evolved into the leading cause of mortality worldwide [7]. It is characterized by plaque accumulation in the medium and large arteries, leading to luminal narrowing and an increased risk of stroke and other cardiovascular events. Early detection and proper treatment of atherosclerosis are crucial for preventing the progression of cardiovascular disease. However, traditional angiographic techniques often underestimate the actual burden of the disease due to the outward remodeling of the arterial wall.

Black blood vessel wall imaging (BB-VWI) with Magnetic Resonance Imaging (MRI) has emerged as an effective method for visualizing and characterizing normal and diseased arteries, as well as atherosclerotic lesions, without the use of ionizing radiation or contrast media [81, 82]. However, an accurate assessment of the thickness of the carotid artery wall, a critical aspect of atherosclerosis management, remains challenging due to the laborious and expertise-dependent nature of manual segmentation, which can exhibit high intra- and inter-observer variability [83]. Furthermore, automatic segmentation techniques face difficulties in handling complex atherosclerotic lesions and intricate arterial geometries. In contrast to other medical segmentation tasks, where pinpointing the area of interest for segmentation is straightforward, the vessel wall generally has a thickness of 1 millimeter and constitutes less than 0.1% of the total image area. Moreover, the dimensions of the arteries might drastically fluctuate among slices.

In this research work, we address these challenges by proposing a fully automated deep learning method to segment the arterial lumen and vessel wall in MRI. Our approach leverages the advantages of BB-VWI while overcoming the limitations of manual and existing automatic segmentation techniques. We demonstrate the effec-

tiveness of our method in accurately and efficiently segmenting the carotid artery wall, even in cases with complex atherosclerotic lesions and intricate arterial geometries. By streamlining the segmentation process and reducing the burden on healthcare professionals, our proposed technique has the potential to significantly enhance the diagnosis and management of atherosclerosis, ultimately leading to improved patient outcomes.

The developed method is based on EfficientNet-UNet++ and comprises a multilevel automatic process that identifies the vessel wall and performs a segmentation. We further enhance precision through a slice concatenation strategy, which is particularly useful in instances of arterial bifurcations and plaque presence, where artery morphology complicates segmentation. Upon evaluation on a publicly accessible dataset, our method achieved state-of-the-art results, underscoring its effectiveness for vessel wall measurement and advancing atherosclerosis monitoring.

This research work contributes to the field of medical image analysis by providing a more accurate and efficient approach to segmenting the vessel wall in 3D MRI of the carotid artery, which is crucial for monitoring patients with atherosclerosis. More specifically, the main contributions of this paper are:

1. A fully automatic deep learning-based approach for the automatic segmentation of the vessel wall in 3D MRI of the carotid artery, consisting of a multilevel process that first selects the area containing the non-black pixels, then identifies the location of the vessel wall, and finally performs a precise segmentation on the upscaled ROI.

2. A slice concatenation strategy that combines information from adjacent slices of the image to improve segmentation accuracy near the bifurcation of the artery, providing local context.

3. An analysis of the impact of input resolution and local contextual information on the final segmentation of the model, exploring how these factors influence segmentation accuracy and providing insights for optimizing the performance of our approach.

4. An evaluation of the proposed method on a publicly available dataset, with a comparison of the results with state-of-the-art methods, showcasing our approach's effectiveness and potential value in clinical applications.

4.2 Related work

The field of vessel wall segmentation has seen the development of numerous semi-automated and automated strategies. The first approaches were based on classical methods, such as active contour models [84, 85] or active shape models [86]. More recently, Arias-Lorza *et al.* [87] developed a segmentation method that employs an Optimal Surface Graph (OSG) cut approach, aiming to maximize regional probabilities enclosed by coupled surfaces. Other approaches have explored Hough circle detection to identify arterial centers, assuming a circular vessel morphology [88]. Although these methods reduce manual steps and exhibit reasonable agreement for images featuring high vessel wall contrast, semi-automated strategies still require human intervention for tasks such as contour detection [84, 85] and seed point initialization [86, 87].

Deep learning-based methods have recently demonstrated remarkable performance in this field. Convolutional auto-encoders have been shown to attain a high level of agreement with manual contours [89]. However, some significant challenges hinder the effective application of deep learning algorithms, including: (1) the difficulty of automatically recognizing the target artery amidst multiple arteries, (2) the presence of both healthy and unhealthy vessel walls, and (3) the suboptimal use of information from adjacent slices to improve segmentation results. In [90], a tracklet refinement-based localization approach was developed to accurately and robustly determine the center of the lumen of arteries along image slices, thereby providing regions of interest for further segmentation of the vessel wall. They proposed transforming the ring-shaped vessel wall into a polar coordinate system to ensure continuity and precision of the vessel wall boundaries.

Building on this, Alblas *et al.* [91] proposed a multitask regression approach in a polar coordinate system for the segmentation of the vessel wall in black blood MRI scans, using convolutional neural networks (CNN). This approach ensured ring-shaped vessel walls. They also identified a problem-specific training data augmentation technique that significantly affects segmentation performance. This technique involves simulating potential inaccuracies in centerline localization by randomly sampling the center point of the polar image within a small radius of the center of mass. This process generates several slightly different polar images, which are then used to train the CNN, thereby improving the robustness of the method.

Within the scope of the CarOtid vessel wall SegMentation and atherosclerOsis diagnosiS (COSMOS) 2022 challenge [14], a total of 331 participants from 40 countries spent three months developing automatic segmentation methods for both easy/normal

arteries and arteries with carotid plaques and/or near carotid bifurcations. They worked with 50 training cases and 25 testing cases, each containing 432 axial slices. Both training and test datasets are available for academic purposes [92, 93]. The best-performing team, Hu *et al.* [94], presented a fully automated 3D segmentation approach. In their label-propagation (LP) pipeline, nearest-neighbor interpolation first converts the sparse 2-D annotations into a rough 3-D mask; a nnU-Net [95] is trained on these volumes, used to create pseudo-labels for the remaining unlabeled slices, and then re-trained on the full scan. Despite its effectiveness, the method inherits the high memory footprint of full-volume nnU-Net inference and can propagate early errors when plaque morphology changes abruptly. nnU-Net remains the *de facto* baseline for carotid segmentation because it self-configures patch size, architecture depth, and data augmentation, providing competitive performance with minimal manual tuning across datasets.

More recently, Li *et al.* introduced DBF-UNet [96], a two-stage framework that (1) builds continuous center-lines, (2) generates refined pseudo-labels with SAM-Med2D prompts, and (3) trains a lightweight dense-bidirectional UNet. DBF-UNet reports DICE scores of 0.8608 on the COSMOS dataset while using ≈ 2.8M parameters—underscoring how accurate results can emerge from compact models once high-quality pseudo-labels are available.

The release of the Segment-Anything Model (SAM) [97] sparked interest in zero-shot medical segmentation. Nonetheless, the domain gap between natural and MRI is substantial. SAM-Med2D [98] fine-tuned on a 2-D medical corpus, while Li *et al.* exploited SAM-Med2D inside their pseudo-label generation pipeline (the S-RPL strategy) to correct errors near bifurcation [96].

4.3 Methodology

In MRI scans, a large portion of black pixels in each transversal slice lacks informative content. Hence, our initial focus was directed toward the regions of non-black pixels within each slice. Despite cropping the original image, the area under consideration for segmentation remains relatively small. This becomes a more pressing issue as the algorithm navigates more complex scenarios, such as atherosclerotic lesions and arterial bifurcations, where finer details become pivotal. To address these challenges, we propose a novel method for 3D CA MRI segmentation, which is depicted in Figure 4.1.

The algorithm consists of the following steps. First, the central information-

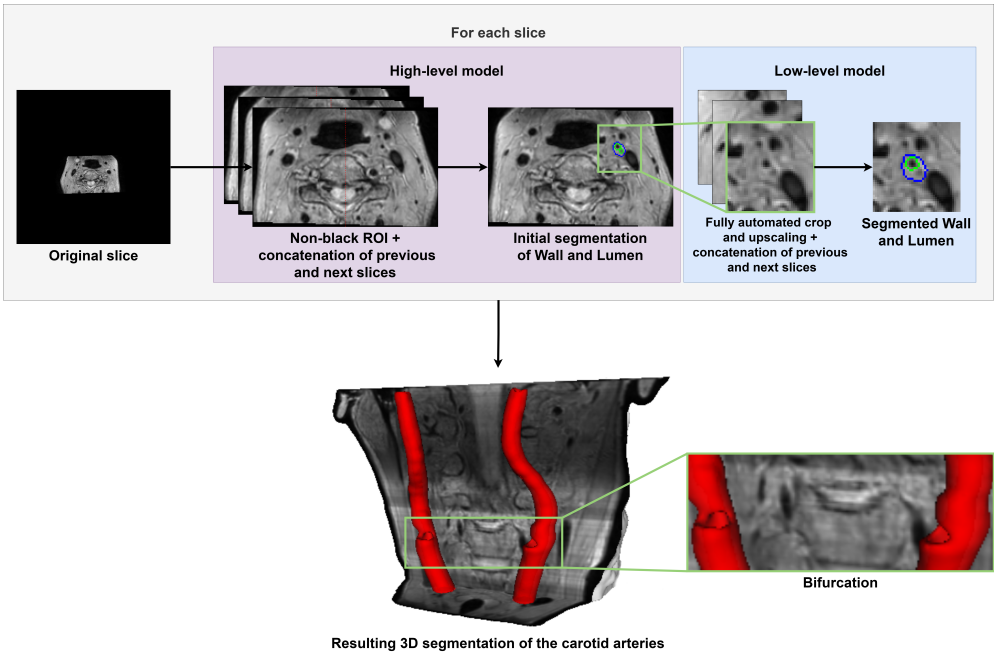

Figure 4.1: Overview of the proposed segmentation pipeline. For each MRI slice, the non-black region of interest is extracted and concatenated with adjacent slices. The concatenated input is split horizontally and processed sequentially by the high-level model to produce an initial segmentation of both arteries. A cropped region around this result is then upscaled and combined with adjacent slices to form the input to the low-level model, yielding refined lumen and wall contours. Repeating this process across slices produces a 3D segmentation of both arteries (bottom), spanning from the upper neck to the lower head. The visible gap in the lower third corresponds to the carotid bifurcation, which is correctly segmented.

containing area is extracted for each slice by removing all black pixels from the slice's borders. The extraction of non-black data is executed through a two-step process. Firstly, the Otsu [99] thresholding technique is applied, iteratively selecting the threshold to maximize the variance between classes of pixels for each slice, effectively separating non-black information from the background. Following this, contour detection is applied to the thresholded binary image to identify the boundaries of the isolated objects, and the contour with the largest area is selected.

Subsequently, an initial segmentation is performed using a model based on UNet++ [100] with EfficientNet B4 [42] as the backbone (high-level model). We obtain the first segmentation mask and use it to determine the location of the region of interest (ROI). Notably, the distinction between the right and left arteries is made based on their position within the image, with both arteries subsequently undergoing identical processing procedures. We then automatically crop their detected regions, leaving a fifteen-pixel margin, and enlarge them to 256×256 pixels. We join the previous and subsequent slices, leaving a $3 \times 256 \times 256$ input matrix. The reasoning behind this margin was to provide sufficient context for correctly selecting the right vessel in cases where the high-level model fails, as well as to detect unusual morphologies. The dimensions of the bounding boxes containing the GT segmentation annotations for the training set are mean height and width values of 15.200 ± 2.642 and 16.103 ± 3.361 pixels, respectively. Then, a second segmentation is performed with the low-level model to obtain the precisely segmented wall and lumen.

4.3.1 Segmentation architecture

The low-level and high-level models share the same architecture but are trained separately. The output is the area between the wall and the lumen. Figure 4.2 describes each of the building blocks of both the encoder and the decoder. Our methodology combines EfficientNet B4, pre-trained on ImageNet [43], and UNet++, leveraging the former's encoding abilities with the latter's decoding expertise to establish an architecture that effectively processes hierarchical features while maintaining precise segmentation boundaries.

The fundamental advantage of EfficientNet compared to conventional convolutional neural networks is its inherently compound scaling method, which simultaneously scales the model's depth, width, and resolution. This substantially reduces the computational cost and memory consumption without compromising the model's performance, making it an ideal candidate for an encoder. Specifically, we replaced the down-sampling

component of the U-Net++ with a pre-trained EfficientNet B4, while the up-sampling maintains the original U-Net++ architecture. The EfficientNet family comprises a set of networks generated through neural architecture search, demonstrating exceptional accuracy while maintaining significantly smaller sizes and faster speeds compared to preceding models. EfficientNet B4 represents one of the proposed variants, striking an optimal balance between the number of parameters and performance efficacy. It achieves 96.4% top-5 accuracy on the ImageNet validation set, requiring only 19 million parameters. The EfficientNet family of networks spans from B0 to B7, with each variant composed of the same building modules (Modules 1 to 4), as shown in Figure 4.2. These modules are assembled into sub-blocks numbered 0 to 3, which are then grouped into seven blocks with varying depths. The primary distinction among these EfficientNet versions stems from the depth of these blocks, with B0 being the lightest and least complex variant, and B7 possessing the highest depth, denoting the most intricate structure within the family of networks. Our study employed the EfficientNet B4 variant due to its balance between complexity and computational demand. Rather than employing depthwise convolution, EfficientNet B4 utilizes a mobile inverted bottleneck with squeeze-excitation optimization, giving it an edge in terms of computational efficiency. Additionally, skip connections from multiple layers (sub-block 0, end of Block 2, Block 3, Block 5, and Block 7) ensure hierarchical features can be learned effectively, capturing both low and high-level features within the network.

The interconnection between EfficientNet and UNet++ begins when the output of Block 7 from the encoder is first upsampled and then concatenated with the output of Block 5. Post concatenation, the result is transferred into sub-block 4, defined by a combination of 2D convolutions, batch normalization [40], and ReLU [101] activation functions in Module 5 —as depicted in Figure 4.2. The same process recursively continues with outputs from different Blocks (5 with 3, 3 with 2, and 2 with the output of the Sub-block 0). The UNet++ decoder, an upgraded variant of U-Net [102], features a deeply supervised encoder-decoder network connected by a series of nested, dense skip pathways. The reasoning behind choosing UNet++ lies in its ability to reduce the semantic gap between the encoder's and decoder's feature maps by enhancing the fine-grained predictions and accurate segmentation boundaries. Additionally, the architecture is designed to capture long-range dependencies and contextual information, thereby optimizing object localization and identification tasks.

The resultant outputs from each level's sub-block 4 undergo concatenation with the output from sub-block 4 of its preceding level, initiating and creating a fully connected upsampling network. This cascade of upsampling continues until the output dimensions

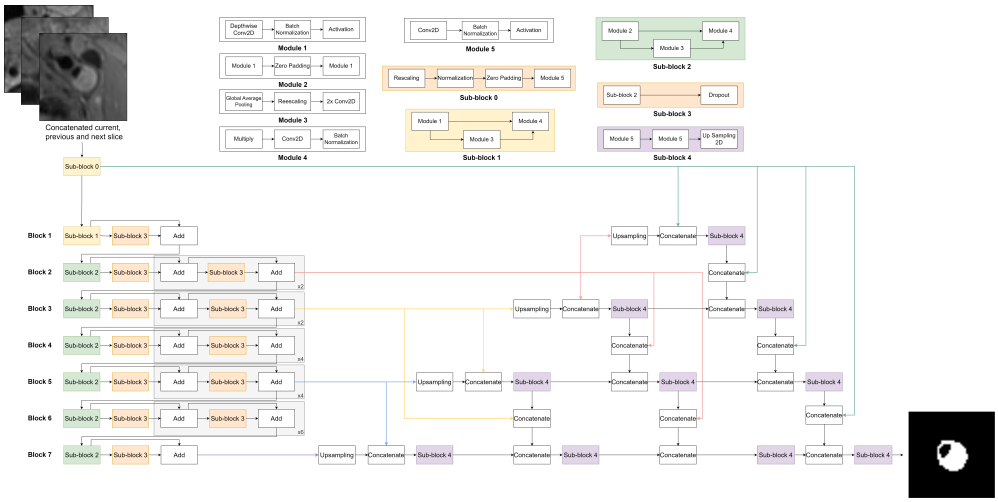

Figure 4.2: Architecture of the proposed segmentation model, consisting of an EfficientNet B4 as encoder and UNet++ as decoder. Skip connections at multiple levels transfer information between encoder and decoder, thus facilitating the re-use of multi-scale features and improving segmentation accuracy.

match the original image resolution.

The chosen loss function was Focal Loss [45], which is defined as follows:

$$\mathrm{FL}\left(p_{\mathrm{t}}\right) = -\alpha_{\mathrm{t}}\left(1-p_{\mathrm{t}}\right)^{\gamma}\log\left(p_{\mathrm{t}}\right) \tag{4.1}$$

$$p_{\mathrm{t}} = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{4.2}$$

where α is a weighting factor, γ is the focusing parameter, and $p \in [0,1]$ is the model's estimated probability for the class with label y.

4.4 Experimental setup

This section presents the dataset utilized for evaluation purposes and the implementation details of the proposed fully automated methodology. Additionally, it includes a description of the experiments conducted and a discussion of the results obtained, including a comparative analysis with the state-of-the-art results on the COSMOS dataset.

4.4.1 Dataset

The proposed method was evaluated using the cohort from the COSMOS challenge [14], which contains 2718 annotated slices from 75 patients. The challenge cohort was derived from clinical data gathered at Renji Hospital, School of Medicine, Shanghai Jiao Tong University (Shanghai, China).

All data were obtained using a 3T Philips MRI scanner equipped with a standard 8-channel carotid coil and a 16-channel head coil, employing the 3D VISTA sequence. The acquisition parameters were as follows: time to echo = 20 ms, repetition time = 800 ms, resolution = $0.6 \times 0.6 \times 0.6$ mm, rows = 432 pixels, and columns = 432 pixels.

Each sample within the dataset consists of an axial resliced 3D image volume, with typical spatial dimensions of 432×432 pixels. Lumen and outer wall contours were drawn by three experienced vessel wall reviewers using the Computer-Aided System for Cardiovascular Disease Evaluation (CASCADE) [103] tool. The arteries analyzed include the internal carotid artery (ICA) and the common carotid artery (CCA). The dataset is divided into two partitions, with 50 cases for training (1825 annotated slices) and 25 (893 annotated slices) for testing, each containing 432 axial slices. Neighboring slices were used to aid in decision-making when the image quality of the current slice was not optimal. Every patient has at least one slice with plaque, with a mean of 12.7500 ± 7.1209 and 12.4545 ± 7.8724 slices with plaque per patient for the training and test sets, respectively. Slices containing plaque or found in the bifurcation zone are identified by a binary variable, which remains unused during the training and testing stages. Only slices containing the carotid artery were annotated in the COSMOS dataset, resulting in non-contiguous ground truth labels along the volume.

Three medically trained reviewers prepared the challenge. Two junior reviewers with more than three years of professional experience in vessel wall research and annotation used the same standard to annotate the vessel wall boundaries and lesion types. Then, a senior reviewer with more than 15 years of professional experience peer-reviewed all the annotations and reached a consensus among all the reviewers.

4.4.2 Implementation details

The proposed framework has been implemented in PyTorch [104], with the source code to be publicly released (upon paper acceptance)[1]. Our approach features a single architecture, described in Section 4.3, that is trained using two distinct datasets: (1)

[1] `https://github.com/gagolucasm/BB-MRI-Carotid-Artery-Segmentation`

three concurrently concatenated slices with the black pixel areas removed, resulting in a high-level model, and (2) three concurrently concatenated slices centered around the ground truth (GT) segmentation, incorporating an additional 15 pixels of contextual information on each side, resulting in a low-level model.

Unlike Hu *et al.*, who interpolated between annotated slices to generate continuous 3D labels for training [94], our approach trained directly on the provided annotations without synthesizing additional labels. We opted for a fully supervised strategy, using only the given ground truth slices.

Each MRI slice contains a large black background with only a small central region holding anatomical information. We therefore first isolate the non-black region of each slice using Otsu's thresholding [99] and contour detection. This concentrates subsequent processing on the area containing the neck and arteries, eliminating meaningless background pixels. Out of slices that are 432×432 pixels before removing the black area, the resulting images have mean height and width values of 106.909 ± 0.839 and 216.628 ± 4.904 pixels, respectively. This preprocessing step is computationally lightweight and deterministic. We found that a simple threshold-based cropping is nearly instantaneous and error-proof for black background removal, negating the need for an extra trained model. In practice, this step added negligible overhead to the testing process.

Each slice was concatenated with its adjacent slices for the high-level model training, forming a triplet consisting of the previous, current, and next slices. The resulting three-channel images were then horizontally cropped in the middle, generating two halves, each containing one of the arteries. The same model was trained to process both the right and left sides by applying random horizontal flips. The primary motivation behind this approach was to simplify the training process, enabling the model to focus on detecting a single object of interest per image.

For the low-level model, we cropped the bounding box containing the GT segmentation with a 15-pixel area on all sides, selected as the mean of the GT bounding box resolution. The next and previous slices, which are cropped around the same area, were concatenated to provide local 3D context.

We employed various data augmentation strategies to prevent overfitting resulting from the use of a relatively small dataset, in addition to the aforementioned random flips. These techniques encompass shifts, scaling, rotations, Gaussian noise addition, perspective transformations, optical distortions, grid distortion, histogram equalization, random gamma correction, blur, motion blur, and random alterations to brightness, contrast, and saturation. Random flips were performed horizontally and vertically,

making the model rotationally invariant.

We extracted 24% of the training data for validation purposes, randomly selecting 12 patients. We used Adam [49] as the optimizer. The parameters β_1 and β_2, representing the exponential decay rate for the first and second-moment estimates, were set to 0.9 and 0.999, respectively. Initially set to 5e−4, the learning rate was reduced on plateau with a patience of 10 and a factor of 0.1. The models were trained with an input resolution of 256×256 and a batch size of 28. Note that early stopping was used with a patience of 30. For the Focal Loss (eqs. 4.1 and 4.2), the weighting factor α was set to 0.25 and the focusing parameter γ to 2.0.

4.4.3 Performance Measures

Our experimental methodology incorporates both quantitative and qualitative assessment strategies to comprehensively evaluate the performance of our models, providing a nuanced understanding of their strengths and limitations in various scenarios.

Quantitative Measures

For the quantitative analysis, we selected the DICE similarity coefficient as the primary metric to assess our model's effectiveness. The DICE measures the overlap accuracy between predicted and GT binary masks and is formulated as:

$$\text{DICE} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4.3}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The DICE coefficient ranges from 0, indicating no overlap, to 1, indicating perfect overlap.

To provide a comprehensive analysis, we evaluated additional performance metrics:

$$\text{LA}_{\text{diff}} = 1 - \frac{|\text{LA}_m - \text{LA}_s|}{\text{LA}_m} \tag{4.4}$$

$$\text{WA}_{\text{diff}} = 1 - \frac{|\text{WA}_m - \text{WA}_s|}{\text{WA}_m} \tag{4.5}$$

$$\text{NWI}_{\text{diff}} = 1 - \frac{|\text{NWI}_m - \text{NWI}_s|}{\text{NWI}_m} \tag{4.6}$$

where LA, WA, and NWI represent lumen area, wall area, and normalized wall index, respectively. Subscripts m and s denote manual and segmented measurements. The

NWI is defined as:

$$\text{NWI} = \frac{\text{OWA} - \text{LA}}{\text{OWA}} \tag{4.7}$$

where OWA is the outer wall area.

We also computed the Hausdorff distance (HD) to measure the maximum discrepancy between the predicted and GT boundaries:

$$\text{HD}(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\} \tag{4.8}$$

where X and Y are the predicted and GT contours respectively, $d(x, y)$ is the Euclidean distance between points x and y, and sup and inf represent the supremum and infimum operators. The Hausdorff distance measures the maximum discrepancy between two sets of points, capturing the worst-case deviation between the predicted and ground truth boundaries.

To account for size variations, we normalized the HD:

$$\text{NHD}_L = \frac{\text{HD}_L}{r_L}, \quad r_L = \sqrt{\frac{\text{LA}}{\pi}} \tag{4.9}$$

$$\text{NHD}_W = \frac{\text{HD}_W}{r_W}, \quad r_W = \sqrt{\frac{\text{OWA}}{\pi}} \tag{4.10}$$

where NHD_L and NHD_W stand for the normalized HD for lumen and wall, respectively, and r_L and r_W are their corresponding radii. Values are truncated to be non-negative.

Qualitative Analysis

To better understand the model's behavior, especially in response to intricate morphologies, we conducted a qualitative analysis. This involved comparing the contour outputs against the GT representations for both lumen and wall structures. We generated color-coded images where green, red, and blue pixels represent TP, FP, and FN predictions, respectively. These visualizations provide an intuitive depiction of the model's segmentation accuracy for each morphological feature.

4.5 Results

We have conducted an extensive evaluation of the proposed segmentation method through three carefully designed experiments. Each experiment aims to assess the robustness and effectiveness of various aspects of the model.

- **Experiment 1: Multilevel model.** The fundamental aim of this experiment is to evaluate the contribution of the secondary segmentation stage (low-level model) to our proposed method. For this purpose, we analyze the effectiveness of our approach with and without the implementation of the second segmentation phase.

- **Experiment 2: Local contextual information.** The objective of this experiment is to investigate the effects of integrating information from adjacent slices into our model. We examine whether this additional contextual information leads to improvements in overall performance, particularly in cases involving challenging slices near bifurcation points or those containing plaque formations.

- **Experiment 3: Input image resolutions.** This experiment aims to analyze the influence of varying the input image resolution of the second model (low-level model) on the performance of our vessel wall segmentation model. We compare the results obtained from input image resolutions ranging from values close to the original size up to more than eleven times larger, assessing their potential effect on segmentation accuracy and precision.

These experiments collectively provide a comprehensive evaluation of our proposed model, enabling us to identify its strengths and areas for potential improvement in the context of vessel wall segmentation and plaque detection in 3D MRI of the carotid artery.

To facilitate a more detailed analysis, the test dataset was partitioned into three distinct scenarios. The first scenario encompasses the entire test set of 893 slices to evaluate the overall performance. The other two scenarios isolate clinically challenging and morphologically complex regions: a subset of 48 slices located near the arterial bifurcation and another subset of 274 slices that contain atherosclerotic plaque. This stratification allows for a targeted assessment of the model's robustness in areas of significant clinical interest.

4.5.1 Experiment 1: Multilevel model

This experiment aims to evaluate the value of the second segmentation step in our proposed multilevel model. To achieve this, we compare the performance of our method with and without the second segmentation step.

We first train and evaluate our deep learning-based approach using only the first segmentation step (one-step model), which identifies the location of the vessel wall in 3D MRI of the carotid artery. We then train and evaluate our multilevel model, which encompasses both the initial identification of the vessel wall location and the subsequent high-resolution segmentation step. We quantitatively measure the performance of both models using the DICE similarity coefficient.

Table 4.1 reports the DICE scores for our two variants, alongside three strong baselines: nnU-Net [95], the COSMOS-2022 winning method by Hu *et al.* [94], and DBF-UNet [96]. Introducing the multilevel stage raises the DICE on the full test set from 0.8352 to 0.8704; it also surpasses nnU-Net (0.8527), DBF-UNet (0.8608), and the challenge winner (0.8563). Near the bifurcation the score climbs from 0.8132 to 0.8534, and on plaque-containing slices from 0.8561 to 0.8699. The coarse stage never failed to include the target vessel inside the automatically cropped ROI, so the fine stage always refined the correct anatomy. A paired t-test on per-patient DICE values confirms the improvement over the one-step model ($t = -2.12,\ p = 0.040$), underscoring the statistical and practical significance of the multilevel design.

Table 4.1: Experiment 1: DICE similarity coefficients for three test scenarios. Best score per row is shown in bold.

Scenario	nnU-Net [95]	*Hu et al.* [94]	DBF-UNet [96]	One-step	Multilevel⋆
Full test set	0.8527	0.8563	0.8608	0.8352	**0.8704**
Near bifurcation	—	—	—	0.8132	**0.8534**
Near plaque	—	—	—	0.8561	**0.8699**

⋆Our proposal

Figure 4.3 qualitatively compares the segmentation results of the one-step model, column (a); and our proposed multilevel method, column (g). The segmentation results are accurate in most straightforward cases. However, some sections have more complex morphologies, which can lead to difficulties in the segmentation process. These complexities may arise due to various factors, such as lower image quality (slice 4), the presence of plaque (slices 2, 4, and 5), challenges in identifying the relevant artery (slices 1, 2, 3, and 7), or regions near the bifurcation (slice 6). In most slices, the second step is crucial for accurately identifying the vessel's location and providing a

high-quality, artifact-free segmentation. Table 4.3 provides additional metrics.

4.5.2 Experiment 2: Local contextual information

In this experiment, we investigate the impact of incorporating contextual information from the previous and next slices on the performance of our model. We employ EfficientNet B4 as the encoder, which we adapt to handle either two-dimensional inputs (a single slice) or three-dimensional inputs (a slice together with its preceding and succeeding slices, referred to as contextual information).

Table 4.2 displays the variations in the DICE coefficient. As can be observed, the addition of contextual information enhances the results regardless of the test scenario. The improvement is relatively subtle when examining the entire dataset, increasing from 0.8658 to 0.8704. However, a more significant improvement is observed when contextual information is crucial for accurate segmentation, both near the bifurcation, rising from 0.8404 to 0.8534; and near the plaque, increasing from 0.8614 to 0.8699. This observation remains consistent with various input resolutions for the low-level model (see Experiment 3 for more details).

Table 4.2: Experiments 2 and 3: DICE similarity coefficients for three test scenarios. The evaluation covers four input image resolutions and examines performance with and without contextual information. Best scores are shown in bold.

			Input image resolution			
Scenario	#Slices	Input type	64×64	128×128	256×256 ⋆	512×512
Full test set	893	Single slice	0.7395	0.8433	0.8658	0.8474
		Contextual information ⋆	0.7426	0.8475	**0.8704**	0.8593
Near bifurcation	48	Single slice	0.7472	0.8286	0.8404	0.8046
		Contextual information ⋆	0.7523	0.8213	**0.8534**	0.8203
Near plaque	274	Single slice	0.7625	0.8493	0.8614	0.8456
		Contextual information ⋆	0.7701	0.8511	**0.8699**	0.8571

⋆ Our proposal

Table 4.3 presents the results of different metrics, consistently indicating better results using contextual information, improving $\mathrm{WA}_{\mathrm{diff}}$ from 0.0624 to 0.0534, $\mathrm{LA}_{\mathrm{diff}}$ from 0.0846 to 0.0567, $\mathrm{NWI}_{\mathrm{diff}}$ from 0.0689 to 0.0669, NHD_W from 0.1424 to 0.1415 and NHD_L from 0.1796 to 0.1761. We use the values reported by Hu *et al.* [94] as the external baseline, as these specific metrics have not yet been published for DBF-UNet.

Figure 4.3 allows us to compare the segmentation results of the model without contextual information —(b), (d), (f)— and with contextual information —(c), (e), (g). It is evident that in slices containing plaques, as in (2), (4), and (5), the contextual

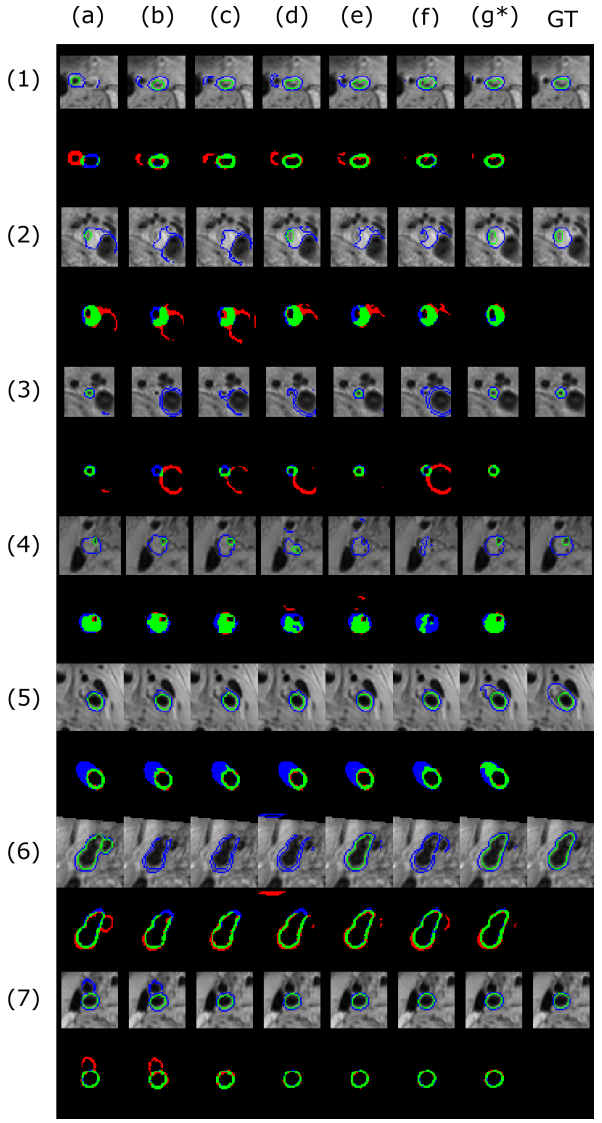

Figure 4.3: Qualitative results on seven representative test slices for all experimental configurations. a) One-step model; b) and c) Multilevel model (input resolution of 64×64), without and with spatial information; d) and e) Multilevel model (input resolution 128×128), without and with spatial information. f) and g) Multilevel model (input resolution of 256×256), without and with spatial information (* our proposal). Each slice is shown in two rows: the top row displays the segmentation outlines (wall in blue, lumen in green), and the bottom illustrates shows the TP pixels in green, FP in red, and FN in blue.

Table 4.3: Additional metrics for every analyzed configuration of input resolutions and methods. Best scores are shown in bold.

Method	Input Size	WA_{diff}	LA_{diff}	NWI_{diff}	NHD_W	NHD_L
Hu *et al.* [94]	190x190	0.0799 ± 0.0268	0.0766 ± 0.0319	0.0704 ± 0.0330	0.1455 ± 0.0301	0.1957 ± 0.0491
One-step model	256 × 256	0.0652 ± 0.1650	0.0730 ± 0.4151	0.0659 ± 0.0600	0.1535 ± 0.1392	0.1855 ± 0.2293
Single slice	64 × 64	0.0636 ± 0.1199	0.0753 ± 0.3505	0.0777 ± 0.0654	0.1975 ± 0.1004	0.2281 ± 0.1113
	128 × 128	0.0598 ± 0.1299	0.0659 ± 0.3651	0.0735 ± 0.0649	0.1501 ± 0.0983	0.1750 ± 0.0919
	256 × 256	0.0624 ± 0.2027	0.0846 ± 0.5436	0.0689 ± 0.0594	0.1424 ± 0.1164	0.1796 ± 0.1987
	512 × 512	0.0667 ± 0.1800	0.0828 ± 0.3931	0.0712 ± 0.0681	0.1619 ± 0.1866	0.2520 ± 0.8611
Contextual information	64 × 64	0.0609 ± 0.1163	0.0763 ± 0.3747	0.0757 ± 0.0620	0.1935 ± 0.1043	0.2274 ± 0.1111
	128 × 128	0.0606 ± 0.1329	0.0676 ± 0.3596	0.0786 ± 0.0682	0.1462 ± 0.0991	0.1805 ± 0.1080
	256 × 256*	**0.0534 ± 0.0660**	**0.0567 ± 0.0644**	**0.0669 ± 0.0598**	**0.1415 ± 0.1016**	**0.1761 ± 0.1571**
	512 × 512	0.0688 ± 0.2111	0.0868 ± 0.5029	0.0680 ± 0.0604	0.1554 ± 0.1481	0.1902 ± 0.3568

* Our proposal

information contributes to an improvement in prediction quality, enabling better delineation of the wall.

4.5.3 Experiment 3: Input image resolutions

This section analyzes the effect of input image resolution on the efficacy of our proposed model for segmenting vessel walls in 3D MRI scans of the carotid artery. We conduct a comprehensive performance analysis of our model by evaluating it at different input resolutions, including those close to the original resolution (64×64 pixels), nearly three times the original resolution (128×128 pixels), approximately six times the original resolution (256×256 pixels), and finally, significantly higher than the input resolution (512×512 pixels). Note that the cropped area of interest around the segmented vessel before upsampling is approximately 45×45 pixels.

We initiate the experiment by training and evaluating our model using the original input resolution of 64×64 pixels and gradually doubling it until we reach 512×512 pixels. Although we are not adding new information, we are enhancing the model's ability to discern finer details in the images, which may lead to improved segmentation accuracy.

Table 4.2 consistently shows an improvement in the DICE coefficient as the input resolution increases from 64×64 to 256×256, reaching a plateau. The most significant enhancement occurs when the input resolution transitions from 64×64 to 128×128, with the coefficient rising from 0.772 to 0.8443. The metric further increases to 0.8697 when the input resolution is 256×256, and then decreases slightly to 0.8625 with an input resolution of 512×512. This trend is also evident in the analysis of the DICE coefficient near the bifurcation and plaque areas. When comparing input resolutions, the DICE coefficient elevates from 0.7595 to 0.8103 and eventually to 0.8437 with a

256×256 input resolution for the bifurcation area, decreasing to 0.8322 with an input resolution of 512×512. A similar pattern is noted near the plaque area, with values rising from 0.7880 to 0.8470 and ultimately 0.8674 with 256×256, and later decreasing to 0.8651 with an input of 512×512 pixels. A similar conclusion can be drawn about the changes in the DICE coefficient in the models without contextual information, where, in every case, performance improves until an input resolution of 256×256 pixels is reached. Table 4.3 shows the impact of input size on additional metrics, and they follow a similar pattern. In every case, the best results are obtained with an input resolution of 256×256 pixels.

Finally, Figure 4.3 illustrates the qualitative impact of varying input resolutions. Columns (b) and (c) have an input resolution of 64×64, (d) and (e) have an input resolution of 128×128, and (f) and (g) have an input resolution of 256×256. Visually, the results consistently improve as we move from left to right across the figure. In some instances, such as slices (1), (2), (3), and (7), lower-resolution models face challenges in correctly identifying the right vessel. Furthermore, in most cases, the number of artifacts observed decreases as the input resolution increases. Note that the results with an input resolution of 512×512 are not shown in Figure 4.3, as they were inferior to those obtained with 256×256, which is the resolution used in our proposed method.

4.5.4 Clinical discussion

The introduction of MRI has substantially transformed the field of cardiovascular research, providing a non-invasive means to explore the pathological features of arterial walls in large populations [105]. However, the full exploitation of these advanced imaging resources depends on the availability of sophisticated analytical tools, such as our proposed segmentation model for the carotid artery using black blood MRI.

A major clinical advantage of our proposal is its ability to support quantitative assessments of the structures of blood vessel walls. The accurate delineation of the vessel lumen and outer wall contours is a fundamental requirement for these assessments, which are crucial for understanding the mechanisms underlying atherosclerotic diseases and evaluating the risk of cardiovascular events [90]. Our fully automatic segmentation method offers a reliable and efficient approach to obtaining these contours, thereby improving the accuracy and precision of the measurements. Since increased carotid intima-media thickness (CIMT) reflects atherosclerosis-related morphological changes and can predict future cardiovascular events [106], our high-quality 3D segmentations enable the extraction of clinically relevant metrics, such as mean and maximum CIMT.

Previous research indicates that mean CIMT is associated with cardiovascular risk factors [13], while maximum CIMT is useful for the detection of atherosclerotic plaques [8]. Extracting these values, particularly in the bifurcation area or in the presence of plaque, often poses challenges. However, our proposed method exhibits robust performance, especially when incorporating local contextual information (see Section 4.5.2).

As demonstrated in Table 4.1, our approach exhibits superior performance compared to the current state-of-the-art methods on the COSMOS dataset [94], while preserving the efficacy on plaque slices. Further, Figure 4.3 provides a visual representation of more intricate scenarios, illustrating the precision of our method in generating accurate results on plaque-containing slices, as well as slices in the bifurcation area.

Moreover, the model addresses a significant obstacle in the utilization of carotid MRI to characterize plaque morphology: the variability among readers. This variability has been recognized as the most substantial constraint for MRI measurements of plaque components [83, 107]. By offering consistent and objective measurements, our proposal can substantially mitigate reader variability, thereby enhancing the dependability of plaque characterization.

4.6 Conclusions

In attempting to solve the complex task of segmenting vessel walls in 3D MRI of the carotid artery, we propose a method that yields significant results and outperforms state-of-the-art approaches. Our proposed method is based on a segmentation network that utilizes EfficientNet B4 as the encoder and UNet++ as the decoder. It includes local 3D contextual information used first to produce a high-level segmentation and then to refine it, focusing only on the region of interest. The multilevel process and the slice concatenation strategy have enhanced the segmentation accuracy and robustness, especially in the most complex scenarios (presence of atherosclerotic plaques and vessel bifurcations).

The proposed strategy achieves leading performance on the COSMOS 2022 dataset. It effectively addresses common challenges in medical imaging, including the small size of the target area compared to the overall image and its irregular shape with high variability. Our segmentation approach was benchmarked against contemporary, state-of-the-art methods, and the results demonstrate a clear improvement over all previously reported performances.

We conducted systematic experiments to assess the influence of contextual informa-

tion on our model's performance. These tests demonstrated significant performance enhancements across the evaluated scenarios. Moreover, our investigations into the impact of the model's input resolution for assessing the cropped region of interest revealed a consistently improved performance when the input resolution was increased from 64×64 to 256×256 pixels. Additionally, we examined the impact of the multilevel approach on the model's effectiveness, noting a noticeable improvement in the DICE coefficient upon implementing this strategy.

A key advantage of our algorithm is its resource efficiency, primarily due to the incorporation of a pre-trained encoder from the EfficientNet family and the use of only three slices per prediction. By utilizing local 3D contextual information instead of the complete 3D MRI, we found that we can maximize computational resources while maintaining high-quality output. This strategy also helps mitigate the risk of overfitting and makes the model lighter. It should be noted that the fine segmentation stage (low-level model) is predicated on a reasonably accurate initial localization of the vessel by the high-level model. If the high-level model's output were grossly incorrect, the low-level model would refine the wrong region. In practice, we observed that the high-level EfficientNet-UNet++ successfully identified the correct ROI in all training, validation, and test slices, including challenging cases.

In summary, our segmentation approach offers a promising contribution to the field. Its strengths in handling complex scenarios, efficient resource use, and overall robust performance make it a valuable tool in developing vessel wall segmentation methods. With further refinement and broader testing, it will provide significant benefits to the medical and scientific communities.

Future research may consider exploring additional post-processing techniques, such as integrating anatomical priors or utilizing information from concurrent slices, which could also enhance prediction refinement. Further experiments could be conducted to optimize the amount of 3D information required to address every possible scenario, considering the optimal number of previous and subsequent slices for each prediction. Finally, testing the proposed method with different datasets and imaging modalities would further enhance its validity and applicability.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (MICINN) under Grant PID2022-136436NB-I00 and the Agency for Management of University and Research Grants of Catalonia (AGAUR) under Grant 2021-SGR-01104.

The Agency for Science, Business Competitiveness, and Innovation of the Principality of Asturias in Spain (SEKUENS) is also acknowledged for funding through the project GRU-GIC-24-018.

Conflict of interest statement

The authors report no potential competing interests.

Chapter 5

Quality-Aware Segmentation

Bridging the Quality Gap: Robust Colon Wall Segmentation in Noisy Transabdominal Ultrasound

Lucas Gago[1], **Miguel A. Fernández González**[2], **Justin Engelmann**[3], **Beatriz Remeseiro**[4*], **Laura Igual**[1*]

[1]Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona 08007, Spain

[2]R&D Department, Generative Intelligence S.L., Calle Acequia del Real 2, Málaga 29649, Spain

[3]Institute of Ophthalmology, University College London, Gower Street, London WC1E 6BT, UK

[4]Department of Computer Science, Universidad de Oviedo, Campus de Gijón s/n, Gijón 33203, Spain

*These authors jointly supervised this work.

Abstract

Colon wall segmentation in transabdominal ultrasound is challenging due to variations in image quality, speckle noise, and ambiguous boundaries. Existing methods struggle with low-quality images due to their inability to adapt to varying noise levels, poor boundary definition, and reduced contrast in ultrasound imaging, resulting in inconsistent segmentation performance. We present a novel quality-aware segmentation framework that simultaneously predicts image quality and adapts the segmentation

process accordingly. Our approach uses a U-Net architecture with a ConvNeXt encoder backbone, enhanced with a parallel quality prediction branch that serves as a regularization mechanism. Our model learns robust features by explicitly modeling image quality during training. We evaluate our method on the C-TRUS dataset and demonstrate superior performance compared to state-of-the-art approaches, particularly on challenging low-quality images. Our method achieves Dice scores of 0.7780, 0.7025, and 0.5970 for high, medium, and low-quality images, respectively. The proposed quality-aware segmentation framework represents a significant step toward clinically viable automated colon wall segmentation systems.

Keywords: Deep learning, Ultrasound imaging, Colon wall segmentation, Image quality assessment, Medical image analysis

Resum

La segmentació de la paret del còlon en l'ecografia transabdominal és un repte a causa de les variacions en la qualitat de la imatge, el soroll speckle i els límits ambigus. Els mètodes existents tenen dificultats amb les imatges de baixa qualitat a causa de la seva incapacitat per adaptar-se a nivells de soroll variables, la definició deficient dels límits i el contrast reduït en la imatge ecogràfica, cosa que resulta en un rendiment de segmentació inconsistent. Presentem un nou marc de treball (framework) de segmentació sensible a la qualitat que prediu simultàniament la qualitat de la imatge i adapta el procés de segmentació en conseqüència. El nostre enfocament utilitza una arquitectura U-Net amb una xarxa troncal (backbone) codificadora ConvNeXt, millorada amb una branca paral·lela de predicció de qualitat que serveix com a mecanisme de regularització. El nostre model aprèn característiques robustes modelant explícitament la qualitat de la imatge durant l'entrenament. Avaluem el nostre mètode en el conjunt de dades C-TRUS i demostrem un rendiment superior en comparació amb els enfocaments de l'estat de l'art, especialment en imatges difícils de baixa qualitat. El nostre mètode aconsegueix puntuacions Dice de 0,7780, 0,7025 i 0,5970 per a imatges d'alta, mitjana i baixa qualitat, respectivament. El marc de segmentació sensible a la qualitat proposat representa un pas significatiu cap a sistemes automatitzats de segmentació de la paret del còlon clínicament viables.

Paraules clau: Aprenentatge profund (Deep learning), Imatge per ultrasons, Segmentació de la paret del còlon, Avaluació de la qualitat de la imatge, Anàlisi d'imatge

mèdica

5.1 Introduction

Ulcerative colitis is a chronic inflammatory bowel disease affecting millions of people worldwide [11]. Monitoring disease activity is crucial for effective treatment management, with current approaches relying heavily on invasive procedures such as colonoscopy. Transabdominal ultrasound has emerged as a promising non-invasive alternative for assessing the colon wall [6, 108], but requires specialized expertise for interpretation.

Automated segmentation of the colon wall in ultrasound images could significantly improve clinical workflows and diagnostic accessibility; however, this task presents unique challenges. Image quality varies significantly due to differences in patient anatomy, operator skill, and machine settings, markedly affecting segmentation performance [15, 12]. Moreover, ultrasound imaging is inherently affected by speckle noise and artifacts that obscure critical structures and reduce segmentation accuracy [17, 109]. The colon wall is a relatively small structure whose visual features closely resemble those of the surrounding tissues, making it challenging to delineate [110]. Finally, the colon wall may present multiple shapes and appearances across different patients and imaging conditions, further complicating the segmentation process [15].

Previous work by Leenings et al. [12] introduced the C-TRUS dataset and established initial benchmarks for colon wall segmentation, reporting moderate performance across different segmentation architectures. Notably, their analysis revealed substantial inter-observer variability among clinical experts, with an average Dice score of 0.6134, highlighting the inherent difficulty of the task.

Current segmentation approaches fail to account for quality variations in clinical ultrasound images, performing poorly on low-quality samples. We propose a quality-aware segmentation framework that explicitly models image quality as an auxiliary task, employs quality predictions to weight loss contributions, and implements quality-dependent regularization. Our dual-branch architecture, based on U-Net with ConvNeXt encoder, simultaneously performs segmentation and quality assessment, adaptively focusing learning on reliable samples through a quality-weighted loss function. Comprehensive data augmentation strategies address ultrasound-specific

challenges. Results demonstrate significant performance improvements over previous methods, particularly on challenging medium and low-quality images. Our framework advances clinically viable automated colon wall segmentation by robustly handling variable image quality.

5.1.1 Related Work

Deep learning approaches have shown promising results for various anatomical structures in ultrasound imaging. Methods for segmenting the kidneys [111], liver [112], and bladder [113] have been developed with varying degrees of success. These approaches often leverage geometric constraints or shape priors to guide segmentation in the presence of noise and artifacts [114, 115].

The unique challenges of ultrasound imaging, including motion blurring, weak boundaries, acoustic shadows, and speckle noise, have prompted specific architectural adaptations [116, 117]. Traditional CNNs struggle with the low signal-to-noise ratio characteristic of ultrasound, leading to innovations such as attention mechanisms and multi-scale feature extraction. Transformer-based architectures have recently shown promise for medical image segmentation, with models like TMU-Net [118] and HiFormer [119] demonstrating effectiveness on various anatomical structures.

Recent ultrasound segmentation advances further strengthen U-Net backbones for noisy B-mode data using attention and multiscale refinement, including AAU-Net [120], which replaces plain convolutions with a hybrid adaptive attention module combining multi-kernel features and channel/spatial self-attention for breast lesions, reporting consistent gains on BUSI/Dataset B and an external set. NU-Net [121] nests shared-weight U-Nets of different depths around a deeper backbone and introduces multi-step down-sampling short connections, improving Dice/Jaccard on multiple breast datasets and transferring to renal ultrasound with competitive boundary metrics. C-Net [122] cascades a coarse U-Net with a bidirectional global–local guidance module and a residual refinement head to learn mask deltas, outperforming baselines on BUSIS and generalizing to renal ultrasound. These methods emphasize multiscale context, attention, and refinement for lesion-centric tasks, whereas the present framework targets colon wall segmentation and explicitly supervises a quality branch to enable quality-aware training and stratified evaluation, which these works do not report.

MTANet [123] is a joint segmentation–classification framework that shares features and applies attention modules to improve both tasks, demonstrating gains on polyp and skin lesion benchmarks and a private liver ultrasound dataset, but it does not

model acquisition image quality nor condition the segmentation objective using quality labels. In contrast, the present framework explicitly supervises per-image acquisition quality and uses quality-weighted objectives and stratified reporting on C-TRUS to make segmentation quality-aware in colon ultrasound, which is a different problem setting than MTANet's diagnosis-oriented multi-task design.

Segmentation of the colon wall in transabdominal ultrasound remains relatively unexplored. Pahl et al. [109] employed Gabor filters for pre-processing ultrasound images to facilitate colon wall thickness measurements. More recently, Leenings et al. [12] established the first comprehensive benchmark using the C-TRUS dataset, evaluating several segmentation architectures including nnU-Net [124], DeepLabv3+ [125], TMU-Net [118], SegNext [126], HiFormer [119], and Mask R-CNN [127]. Their analysis revealed the substantial challenge of this task, with inter-observer variability among clinical experts yielding only moderate agreement (Dice score of 0.6134).

Several approaches incorporate image quality into different medical imaging segmentation frameworks. Quality-driven attention mechanisms [117] and uncertainty estimation [128] have been employed to improve robustness. Multi-task frameworks that jointly predict segmentation masks and quality scores have shown improved performance in cardiac MRI and retinal imaging. However, these approaches typically do not explicitly model image quality as a learning target or use it to guide loss weighting during training, particularly in the ultrasound domain.

Multi-task learning has shown promise in medical imaging by leveraging shared representations across related tasks [129]. Auxiliary tasks such as edge detection and distance map regression can enhance primary segmentation performance, acting as regularizers that prevent overfitting [130]. This is particularly valuable in medical imaging, where annotated data is often limited.

Data augmentation plays a crucial role in training robust segmentation models, especially for ultrasound imaging. Beyond traditional geometric transformations, ultrasound-specific augmentations that simulate imaging artifacts and quality variations have shown superior performance [131]. Mixup [132] and its variants provide smooth interpolations between training samples that improve model calibration and generalization.

Our work bridges these areas by introducing a novel quality-aware segmentation framework specifically tailored to the challenges of colon wall segmentation in transabdominal ultrasound.

5.1.2 Contributions

In this paper, we introduce a quality-aware segmentation framework for automated colon wall segmentation in transabdominal ultrasound images. The proposed framework is a dual-branch deep learning system comprised of: (1) a U-Net with ConvNeXt encoder backbone for semantic segmentation and (2) a parallel quality assessment branch for image quality prediction. The segmentation model learns robust features across varying image conditions through quality-weighted loss functions and comprehensive data augmentation specifically tailored to ultrasound imaging challenges. The quality assessment branch simultaneously classifies images into high, medium, or low quality categories, enabling adaptive learning based on expected annotation reliability.

The main contributions of this research are the following:

- A quality-aware segmentation model (QA U-Net) composed of: (1) a semantic segmentation branch based on U-Net with ConvNeXt encoder, achieving state-of-the-art results across all quality levels; and (2) a quality assessment branch that predicts image quality and guides the learning process through quality-weighted loss functions, resulting in Dice scores of 0.7780, 0.7025, and 0.5970 for high, medium, and low-quality images, respectively.

- A comprehensive ablation study demonstrating the effectiveness of each component in our framework, including statistically significant improvements in precision (6.79%, $p = 0.0005$) and high-quality segmentation masks (13.50% relative improvement, $p = 0.0447$), along with a training strategy incorporating mixup and gradient clipping for enhanced generalization on small medical datasets.

- A thorough comparison with six state-of-the-art segmentation methods on the C-TRUS dataset, demonstrating superior performance, particularly on challenging medium and low-quality images, with our model achieving an overall Dice score of 0.7137 that exceeds the average inter-observer agreement among medical experts (0.6134).

The rest of the paper is structured as follows. Section 5.2 details the proposed quality-aware framework, loss functions, and training strategy. Section 5.3 presents experimental results and ablation studies. Section 5.4 discusses clinical implications and limitations. Finally, Section 5.5 concludes with future directions.

5.2 Materials and Methods

5.2.1 Dataset

We utilize the C-TRUS dataset[1] [12], which contains 827 transabdominal ultrasound images (580 × 360 pixels) taken with a Toshiba Xario General LCD, from 13 patients with ulcerative colitis (mean age of 40.25 ± 15.37 years, seven females). The dataset includes expert annotations of the colon wall and subjective image quality ratings: high – 185 images (22.4%), medium – 295 images (35.7%), low – 347 images (41.9%), provided by experienced gastroenterologists. In C-TRUS, image quality was rated subjectively on a three-level scale (low/medium/high) by a single, most experienced annotator to enable quality-stratified analyses and to study the impact of quality on manual and automated segmentation; we present representative examples in Figure 5.1 and use these labels for training and stratified reporting in this study. The colon wall is present in 507 of the 827 images, appearing in various shapes (355 lines, 129 arcs, and 23 variations).

The dataset highlights the considerable challenge of this segmentation task, with inter-observer agreement among experts yielding only a moderate Dice score of 0.6134. Notably, expert disagreement about the presence of the colon wall occurs in 21.85% of cases overall and rises to 34.34% for low-quality images.

5.2.2 Quality-aware U-Net

We propose Quality-aware U-Net (QA U-Net), a dual-branch architecture that simultaneously performs colon wall segmentation and image quality assessment. Our model is built upon U-Net [102], incorporating two key modifications detailed below. Figure 5.2 illustrates the complete architecture of our proposed QA U-Net.

ConvNeXt Encoder: We replace the traditional convolutional encoder with a ConvNeXt [133] backbone, specifically ConvNeXt-Base with 4 stages and block depths of [3, 3, 27, 3] respectively. Each ConvNeXt block contains 3 convolution operations: one depthwise convolution (7×7 kernel) and two pointwise convolutions (1×1 kernel), resulting in 108 total convolutions in the encoder. The feature channels progress as [128, 256, 512, 1024] across the four stages. ConvNeXt modernizes the standard ConvNet architecture by incorporating design principles from Vision Transformers, thereby providing stronger feature extraction capabilities while maintaining the inductive biases that are beneficial for medical image analysis. Notably, our experiments showed a

[1]The dataset is available at `https://github.com/wwu-mmll/c-trus`.

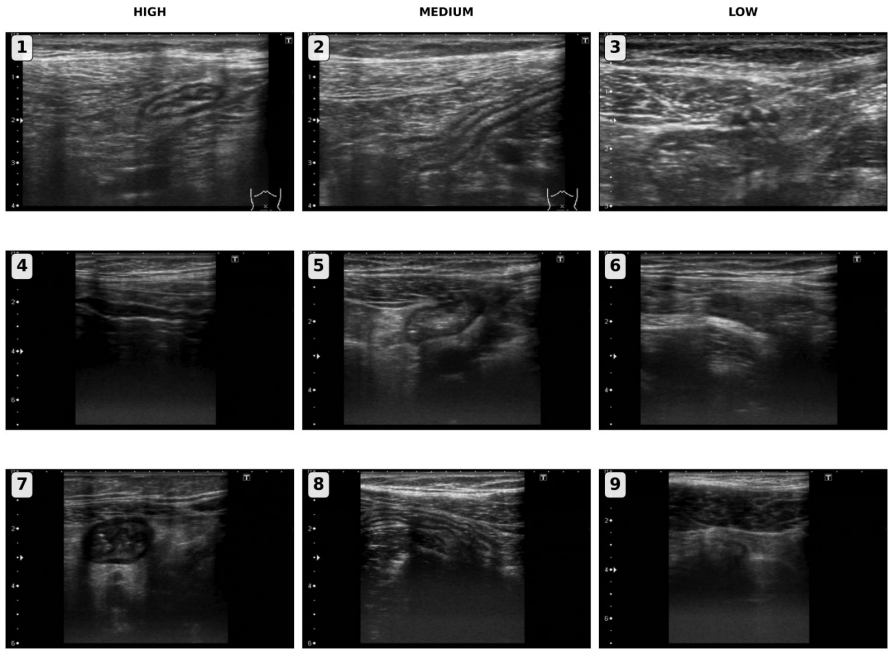

Figure 5.1: Representative ultrasound images from colon segmentation database classified by image quality. (1, 4, 7) High-quality images exhibit clear colonic structure definition with optimal tissue contrast, enabling straightforward segmentation procedures. (2, 5, 8) Medium-quality images present identifiable colonic anatomy but demonstrate reduced contrast that may challenge accurate segmentation boundaries. (3, 6, 9) Low-quality images show poor structural differentiation with significant noise interference, making colonic structure recognition and segmentation procedures difficult or unreliable. Image quality assessment directly impacts segmentation accuracy and clinical diagnostic utility.

significant improvement when training these models compared to classical CNNs.

Dual-Branch Decoder: QA U-Net features two parallel decoder branches. The *Segmentation Branch* follows a standard U-Net decoder structure, consisting of four symmetric decoder stages. Each decoder block contains three convolution operations: one upsampling/transpose convolution and two regular convolutions (3×3 kernel) for feature processing after skip connection concatenation, generating pixel-wise segmentation masks. In parallel, the *Quality Assessment Branch* serves as a lightweight classification head responsible for predicting image quality as high, medium, or low. This branch consists of a global average pooling layer followed by two fully connected layers with ReLU activation and dropout (rate = 0.2). The final layer outputs three logits corresponding to the quality classes.

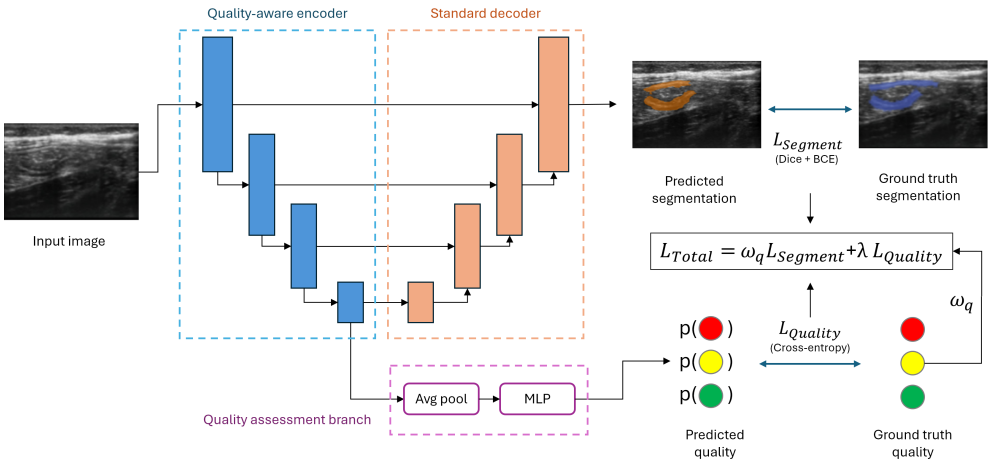

Figure 5.2: Architecture of the proposed Quality-aware U-Net (QA U-Net) for colon wall segmentation in transabdominal ultrasound. The model consists of a ConvNeXt encoder backbone (blue blocks) with four stages of increasing feature channels, followed by a bottleneck layer. From the bottleneck, the encoded features split into two distinct parallel decoder branches: (1) a segmentation decoder branch (orange blocks) with symmetric upsampling blocks connected via skip connections that produces the final segmentation mask, and (2) a quality assessment branch (purple blocks) with global average pooling and fully connected layers that classifies image quality into High/Medium/Low categories. This dual-branch decoder design enables simultaneous segmentation and quality prediction. The model is trained with a quality-weighted total loss that combines segmentation loss (BCE + Dice) and quality prediction loss, where the quality weight adaptively adjusts the contribution based on image quality.

5.2.3 Quality-Weighted Loss Function

We introduce a quality-weighted loss function that adaptively adjusts the contribution of each sample based on its predicted quality:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{quality} \tag{5.1}$$

where $\mathcal{L}_{quality}$ is a categorical cross-entropy loss for the three-class quality prediction, and λ is a weighting factor (set to 0.2 in our experiments).

The segmentation loss $\mathcal{L}_{seg}$ combines binary cross-entropy loss [134] and Dice loss [135], weighted by the ground truth quality label during training:

$$\mathcal{L}_{seg} = \omega_q \cdot (\mathcal{L}_{BCE} + \mathcal{L}_{Dice}) \tag{5.2}$$

where ω_q is a quality-dependent weight:

$$\omega_q = \begin{cases} 1.0 & \text{if quality is high} \\ 0.75 & \text{if quality is medium} \\ 0.5 & \text{if quality is low} \end{cases} \tag{5.3}$$

This weighting scheme directly addresses the substantial annotation unreliability in low-quality images, where expert disagreement regarding colon wall presence reaches 34.34% compared to only 21.85% overall. By reducing the contribution of these uncertain labels to the training signal, we prevent the model from learning from potentially inconsistent ground truth annotations, which improves training stability and convergence in our experiments.

5.2.4 Training Strategy

To enhance training diversity and bolster generalization, we employ a comprehensive suite of augmentations that address both spatial and intensity variations. Our approach includes spatial transformations such as horizontal flips, elastic deformations, random rotations of up to ±15 degrees, scaling within a 0.9–1.1 range, and translations of up to ±10%. All of these transformations simulate the natural variability in imaging conditions. In parallel, we apply intensity adjustments including Gaussian blur (with σ ranging from 0.5 to 1.5), CLAHE (using a clip limit of 2.0), and random brightness and contrast modifications of up to ±20%. To further improve generalization and regularization, we incorporate mixup [132] with $\alpha = 0.2$, where for each training batch, images and their corresponding masks are linearly interpolated with a probability of 0.2. Additionally, gradient clipping [136] is applied with a threshold of 1.0 to stabilize training and prevent exploding gradients. This integrated augmentation and regularization strategy effectively mimics real-world ultrasound variations, ultimately contributing to a more robust and generalizable model.

We train our model using the AdamW optimizer [137] with a learning rate of 5e-5. The network is trained on 512×512×1 input images resized from full-resolution images (580 × 360) with a batch size of 4, while a weight decay of 1e-4 is applied to regularize the learning process. Training is performed for up to 100 epochs, employing early stopping with a patience of 15 epochs.

The code related to the training and evaluation of the models will be released upon acceptance of this paper.

5.3 Experiments and Results

5.3.1 Experimental Setup

Following the protocol established in [12], we employ five-fold cross-validation with patient-wise splits to prevent data leakage, with the folds provided in the dataset.

We evaluate the segmentation performance of QA U-Net using standard metrics: Dice coefficient, Precision, and Recall. Following [12], we also report the percentage of cases with Dice scores ≥ 0.50 and ≥ 0.75.

We reimplemented all baseline methods according to the specifications in [12] to ensure a fair comparison. Since the original code for these baselines was not provided, we relied solely on their detailed instructions. This approach produced consistent results for nnU-Net, DeepLabV3+, HiFormer, TMU-Net, and SegNext but not for Mask R-CNN, where our reproduced outcomes deviated from those reported in the original paper.

All experiments were conducted using PyTorch [104] version 2.6 as our primary deep learning framework. The training process was conducted on an NVIDIA RTX 4090 GPU.

5.3.2 Results

Table 5.1 presents the performance of our quality-aware segmentation model compared to baseline methods across different image quality categories. For the Mask R-CNN model, we report our reproduced results rather than those claimed in the original paper.

Our quality-aware segmentation model consistently outperforms baseline methods across all image quality categories. The performance improvements are particularly significant for medium and low-quality images, demonstrating the effectiveness of our quality-aware approach in handling challenging cases. Figure 5.3 illustrates how our model performs compared with the other analyzed options, showing strong performance in complex scenarios. Statistical significance testing using Welch's t-test (accounting for unequal variances between methods) demonstrates the robustness of our quality-aware approach. Across all images, QA U-Net achieved highly significant improvements in all five metrics compared to Mask R-CNN (Dice, Precision, Dice ≥ 0.50, and Dice ≥ 0.75: all $p < 0.0001$; Recall: $p = 0.0002$). The performance gains varied by image quality level: (i) For high-quality images, where Mask R-CNN already performs well (Dice = 0.759), QA U-Net showed significant improvement only in Precision (0.809 vs 0.764, p

Table 5.1: Performance of different models on various image quality levels (with updated Mask R-CNN values and QA U-Net results). Best results are marked in bold, and Asterisks (*) indicate statistically significant improvements ($p < 0.05$) compared to Mask R-CNN.

Quality level	Model	Dice	Recall	Precision	Dice ≥ 0.50	Dice ≥ 0.75
All images	Mask R-CNN	0.6231	0.7020	0.6508	73.62%	42.32%
	nnU-Net	0.5195	0.6247	0.5119	59.17%	25.25%
	DeepLabV3+	0.5012	0.5884	0.5110	59.96%	16.77%
	HiFormer	0.4596	0.5810	0.4518	50.10%	6.88%
	TMU-Net	0.4201	0.4294	0.5341	44.40%	8.84%
	SegNext	0.4046	0.3725	0.5581	38.46%	0.59%
	QA U-Net	**0.7137***	**0.7534***	**0.7378***	**88.39%***	**52.95%***
High quality images	Mask R-CNN	0.7590	**0.8079**	0.7636	91.18%	67.06%
	nnU-Net	0.6603	0.7318	0.6467	82.25%	46.15%
	DeepLabV3+	0.6236	0.6914	0.6134	79.29%	31.36%
	HiFormer	0.5632	0.7106	0.5191	71.60%	13.02%
	TMU-Net	0.5458	0.5639	0.6284	62.13%	18.93%
	SegNext	0.4841	0.4466	0.6223	53.85%	0.00%
	QA U-Net	**0.7794**	0.7910	**0.8094***	**94.71%**	**71.18%**
Medium quality images	Mask R-CNN	0.5948	0.6806	0.6138	70.61%	35.92%
	nnU-Net	0.4790	0.6055	0.4653	50.61%	18.78%
	DeepLabV3+	0.4703	0.5769	0.4854	55.10%	10.20%
	HiFormer	0.4344	0.5492	0.4400	43.67%	3.67%
	TMU-Net	0.3987	0.4158	0.5199	41.63%	4.90%
	SegNext	0.3831	0.3467	0.5507	33.47%	0.82%
	QA U-Net	**0.7149***	**0.7566***	**0.7308***	**90.20%***	**50.61%***
Low quality images	Mask R-CNN	0.4496	0.5647	0.5418	49.46%	13.98%
	nnU-Net	0.3702	0.4806	0.3897	39.78%	4.30%
	DeepLabV3+	0.3602	0.4314	0.3928	37.63%	7.53%
	HiFormer	0.3400	0.4324	0.3623	28.42%	4.21%
	TMU-Net	0.2517	0.2254	0.4031	20.00%	1.05%
	SegNext	0.3169	0.3060	0.4611	23.66%	1.08%
	QA U-Net	**0.5904***	**0.6765***	**0.6253**	**72.04%***	**25.81%***

= 0.037), with other metrics showing no significant differences; (ii) For medium-quality images, all metrics showed highly significant improvements (Dice, Precision, Dice $\geq$ 0.50, and Dice $\geq$ 0.75: all $p < 0.0001$; Recall: $p = 0.0002$), with Dice improving from 0.595 to 0.715 and the percentage of well-segmented images (Dice $\geq$ 0.75) increasing from 35.9% to 50.6%; (iii) For low-quality images, significant improvements were observed in Dice (0.590 vs 0.450, $p = 0.0002$), Recall (0.677 vs 0.565, $p = 0.011$), Dice $\geq$ 0.50 ($p < 0.0001$), and Dice $\geq$ 0.75 ($p = 0.002$), while Precision improvement approached but did not reach significance ($p = 0.059$).

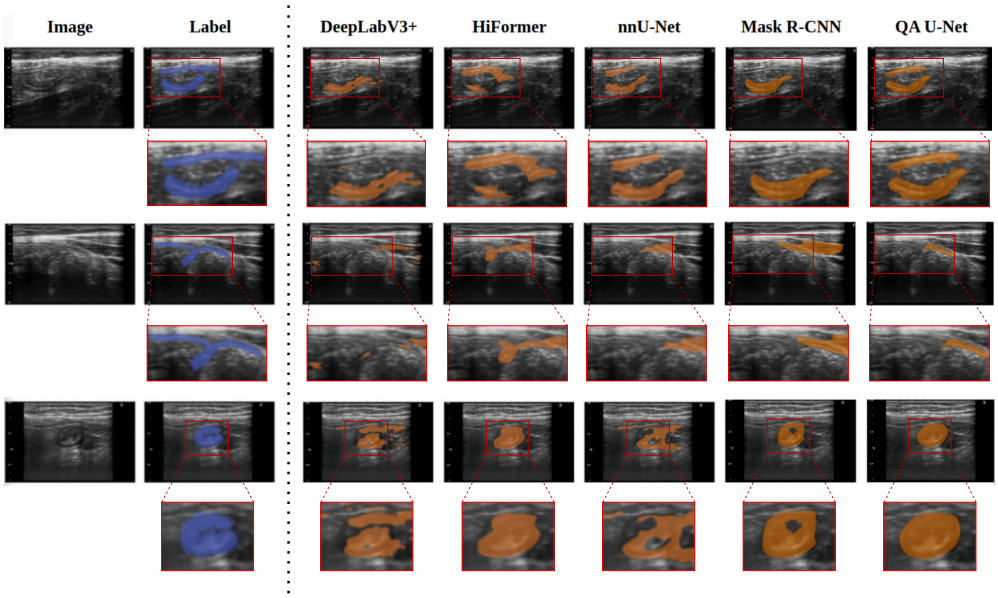

Figure 5.3: Qualitative results on representative images of different quality levels. Our quality-aware approach maintains robust segmentation performance even on low-quality images.

5.3.3 Ablation Study

We conduct ablation experiments to analyze the contribution of individual components in our framework (Table 5.2). Statistical significance tests were performed across all test images using Welch's t-test for continuous metrics and z-test for proportional metrics.

Our statistical analysis reveals that the final model, incorporating all proposed components, exhibits statistically significant improvements in key metrics. The Dice score improved by 3.58% ($p=0.0253$) compared to the base model. The most substantial improvement was observed in precision, with a 6.79% increase ($p=0.0005$) over the base model. Notably, each incremental addition of the quality-weighted loss mechanism provided statistically significant precision improvements ($p=0.0256$ and $p=0.0166$, respectively) compared to prior variants.

While the overall percentage of images with Dice $\geq 50\%$ increased from 85.04% to 88.39%, this improvement was not statistically significant ($p=0.1161$). However, the proportion of high-quality segmentation masks (Dice $\geq 75\%$) increased significantly from 46.65% to 52.95% ($p=0.0447$), representing a 13.50% relative improvement.

Table 5.2: Ablation study results showing the impact of different components in our framework. Asterisks (*) indicate statistically significant improvements ($p < 0.05$) compared to the base model.

Variant	Dice	Precision	Recall	Dice $\geq 50\%$	Dice $\geq 75\%$
Base	0.6890	0.6909	0.7563	85.04%	46.65%
+ Quality Branch	0.7019	0.7074	0.7608	85.63%	48.82%
+ Mixup	0.7040	0.7052	**0.7709**	87.99%	48.82%
+ Quality-Weighted Loss	**0.7137***	**0.7378***	0.7534	**88.39%**	**52.95%***

Interestingly, recall metrics showed no statistically significant changes across model variants, indicating that our improvements primarily enhanced precision without compromising recall. This suggests that the quality-weighted approach effectively reduces false positives while maintaining the model's ability to identify true positives.

To demonstrate that our quality-aware improvements extend beyond the specific ConvNeXt-based architecture, we applied our complete pipeline (quality branch + quality-weighted loss + mixup) to Mask R-CNN, the second-best performing baseline method. This experiment yielded substantial improvements across all quality levels: overall Dice increased from 0.6231 to 0.6472, with particularly notable gains on challenging images where medium-quality Dice improved from 0.5948 to 0.6403 and low-quality Dice increased from 0.4496 to 0.5470. While high-quality performance showed a slight decrease (0.7590 to 0.7186), this trade-off resulted in significantly more balanced performance across quality levels and an overall net improvement.

5.3.4 Design rationale

QA U-Net is cast as a *multi-task learner*: the encoder must carry information that is simultaneously predictive of boundary location and global image quality. Information-bottleneck theory shows that this shared constraint regularises the representation and improves generalization in multi-task settings [138]. Replacing the vanilla U-Net encoder with ConvNeXt introduces large depth-wise kernels and modern training heuristics while retaining the locality bias needed to capture speckle texture; ConvNeXt backbones now match or surpass Transformer variants on several medical-segmentation benchmarks, especially in data-limited or noisy regimes [133, 139, 140]. Because label reliability degrades with image quality, we weight the segmentation loss by the predicted quality — an approximation of heteroscedastic-uncertainty weighting that has proven effective in multi-task optimization [141]. Finally, vicinal-risk minimization

via Mixup smooths decision boundaries and mitigates over-fitting and label noise in medical images [132, 142]. The ablation in Table 5.2 confirms that each component yields an additive improvement in Dice and precision.

5.4 Discussion and Limitations

Our experimental results demonstrate that explicitly modeling image quality within the segmentation framework leads to significant improvements in segmentation performance. The performance gains are most pronounced for challenging medium and low-quality images.

By reducing the penalty for disagreement on low-quality samples, the model learns more reliable features from high-quality examples while still benefiting from the diversity of challenging cases. The combination of comprehensive data augmentation, mixup, and gradient clipping effectively prevents overfitting and improves generalization, which is particularly important given the small dataset size and potential noise in the labels. Notably, our reimplementation of the Mask R-CNN architecture was unable to reproduce the results reported in [12]. This highlights the importance of open-source implementations and reproducibility in medical image analysis research.

Although our quality-aware segmentation framework demonstrates promising performance, several limitations persist. The study utilizes the relatively small C-TRUS dataset from a limited number of patients and a single ultrasound device, raising generalizability concerns. Our model relies on subjective quality labels from a limited number of experts, potentially introducing bias. Additionally, our framework is designed for static images, while ultrasound examinations typically involve dynamic sequences.

5.4.1 Clinical Implications

The performance of our model, with a Dice score of 0.7137, exceeds the average inter-observer agreement among medical experts (0.6134) reported in [12]. This suggests that our automated approach can provide consistent and reliable segmentation results comparable to those of experts.

This study deliberately trains and evaluates segmentation exclusively on images with a visible colon wall, focusing on cases with reliable ground-truth masks given the dataset's reported inter-observer variability and the ambiguity of wall presence in some acquisitions; this aligns the scope with the dataset's segmentation labels and

the benchmark protocol. All baseline methods (nnU-Net, DeepLabV3+, HiFormer, TMU-Net, SegNext, and Mask R-CNN) were trained under the same positive-case protocol to ensure a fair, like-for-like comparison of segmentation performance. Looking ahead, an explicit detection stage to robustly handle studies without a visible colon wall is a planned extension to broaden clinical applicability and support deployment-ready workflows in line with contemporary reporting guidance.

The improved performance on low-quality images is particularly valuable for clinical applications, as it reduces the dependency on perfect acquisition conditions and operator expertise. This could significantly expand the accessibility of colon wall ultrasound assessment for monitoring ulcerative colitis.

5.5 Conclusion

We presented a novel quality-aware segmentation framework for colon wall segmentation in transabdominal ultrasound images. Our approach explicitly models image quality as an auxiliary task and uses it to guide the segmentation process, resulting in significant performance improvements across all quality levels, with notable gains for challenging low-quality images. Notably, our strategy of reducing the influence of uncertain annotations in low-quality images through quality-weighted loss functions demonstrates that handling annotation uncertainty is crucial for robust ultrasound segmentation, enabling the model to learn from reliable labels while maintaining performance on challenging cases.

The proposed framework achieves Dice scores of 0.7780, 0.7025, and 0.5970 for high, medium, and low-quality images, respectively, outperforming existing approaches and exceeding the average inter-observer agreement among medical experts.

Future work should validate the framework on larger, multi-institutional ultrasound datasets, such as DDTI [143], which provides B-mode thyroid images with lesion contours, TI-RADS descriptors, and Bethesda pathology but lacks per-image quality labels, so it is not directly applicable to supervising the quality branch without additional labeling or proxy IQA targets, which we plan to develop as a separate resource. Incorporating objective quality metrics through unsupervised methods could improve reliability, while integrating temporal models could enhance segmentation consistency. Further clinical validation is needed to assess diagnostic accuracy and impact on patient management. Extending the framework to extract clinically relevant parameters would enhance its utility for automated colon wall segmentation in transabdominal ultrasound imaging.

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (MICINN) under Grant PID2022-136436NB-I00 and the Agency for Management of University and Research Grants of Catalonia (AGAUR) under Grant 2021-SGR-01104. The Agency for Science, Business Competitiveness, and Innovation of the Principality of Asturias in Spain (SEKUENS) is also acknowledged for funding through the project GRU-GIC-24-018.

Conflict of interest statement

No potential competing interest was reported by the authors.

Chapter 6

Conclusions and Future Work

This thesis has systematically addressed the translation gap between deep learning research and clinical deployment through a compendium of four interconnected articles. Rather than developing isolated technical solutions, this work has tackled the fundamental barriers that prevent AI systems from achieving reliable, efficient, and robust performance in real clinical environments. This chapter synthesizes the collective insights from this research journey, evaluates the broader implications for medical AI, and charts directions for future work.

6.1 Synthesis of Contributions and Insights

The four research articles in this thesis form a coherent progression, moving from foundational efficiency and clinical integration to high-precision analysis in complex modalities and, finally, to the critical challenge of clinical robustness. Each contribution provides not only a technical advance but also a key insight into building medically viable AI.

From Handcrafted Pipelines to End-to-End Efficiency (Chapter 2) As a foundational contribution, our end-to-end CIMT analysis framework demonstrated that learned representations can entirely replace brittle, domain-specific post-processing while achieving superior performance and a 20× speed improvement. This work establishes a critical principle: for clinical viability, especially in real-time modalities like ultrasound, the objective should be a single and efficient model that provides multiple clinically relevant outputs. This approach solves the dual challenges of workflow integration and computational efficiency, providing the necessary baseline upon which

more sophisticated clinical tools can be built.

From Image Features to Clinical Prognosis (Chapter 3) Building directly on the efficient feature extractor from the first study, our integration of deep features into survival models represents a key methodological bridge between medical image analysis and clinical practice. By demonstrating a 20% improvement in patient risk reclassification, this work provides definitive evidence that learned image representations capture prognostic information that is invisible to the human eye and complementary to traditional clinical risk factors. This insight validates the role of AI not as a replacement for clinical judgment, but as a powerful augmentation tool that can uncover novel biomarkers and lead to more personalized patient care.

Achieving High Precision in Complex 3D Modalities (Chapter 4) This thesis extended its scope beyond 2D ultrasound to address the distinct challenges of high-precision segmentation in 3D black-blood MRI. Our multilevel, context-aware framework for carotid vessel wall segmentation showed that achieving state-of-the-art accuracy in complex, low-contrast anatomies requires specialized architectural innovation. The use of a coarse-to-fine refinement strategy and 3D contextual information proved essential. This contribution underscores a vital insight: while the core principles of deep learning are universal, adapting them to the unique physics, dimensionality, and anatomical complexity of different imaging modalities is critical for achieving expert-level precision.

Confronting Data Variability with Quality-Aware Learning (Chapter 5) Finally, our quality-aware segmentation framework directly confronts what is arguably the most significant and persistent barrier to the deployment of ultrasound AI: performance degradation on variable-quality clinical data. By explicitly modeling image quality within the training process, our framework learns to be robust, significantly improving performance on the medium- and low-quality images that cause standard models to fail. The key insight here is that robustness is not an emergent property but must be explicitly designed for. This approach provides a blueprint for creating AI systems that are reliable and trustworthy in the uncontrolled conditions of real-world clinical practice.

Collective Impact on Medical AI Translation Taken together, these four contributions do more than solve isolated problems. They form a methodological roadmap for developing clinically viable medical AI. This progression —from an efficient end-to-end baseline (RQ1), to clinical prognostic integration (RQ2), to precision in complex modalities (RQ3), and finally to quality-aware robustness (RQ4)— provides a comprehensive framework for taking an AI model from a research concept to a tool that is ready for the rigors of clinical validation and deployment.

6.2 Limitations and Critical Assessment

Despite the significant advances presented, it is essential to critically assess the limitations of this work to understand the boundaries of its immediate clinical impact and to guide future research.

Dataset Generalizability and Scope: While our validation encompasses four distinct clinical applications, the geographic and demographic scope of the datasets remains a primary limitation. The cardiovascular ultrasound studies (Chapters 2 and 3) utilized the REGICOR database, a well-characterized but geographically specific Spanish Mediterranean population [13, 50]. The 3D MRI study (Chapter 4) was conducted on the COSMOS 2022 challenge dataset, which originates from a single Chinese academic hospital. Finally, the colon wall study (Chapter 5) employed the C-TRUS dataset from a German cohort of ulcerative colitis patients [12]. These datasets, while valuable, represent relatively homogeneous populations and standardized acquisition protocols. True clinical robustness requires validation across a wider diversity of equipment manufacturers, operator experience levels, patient demographics, and co-morbidities to ensure the models generalize globally [5]. This critical need for large-scale, multi-center validation is outlined as a primary future research direction in Section 6.3.2.

Prospective Validation and Clinical Outcomes: The studies in this thesis were retrospective in nature. While our cardiovascular risk stratification work (Chapter 3) demonstrated a significant improvement in risk reclassification, using as a foundation the model developed in Chapter 2, this is a statistical surrogate for clinical outcomes. The segmentation accuracy in the other chapters, while state-of-the-art, has not yet been prospectively validated to show that it leads to better treatment decisions or improved patient outcomes. A prospective clinical trial would be the necessary next step to confirm the real-world clinical utility of these frameworks. Therefore, designing and conducting prospective clinical trials represents a crucial, albeit long-term, future step to translate these findings into clinical practice.

Subjectivity in Ground Truth Data: The performance of our models is benchmarked against ground truth annotations that are themselves subject to limitations. For the quality-aware framework (Chapter 5), the quality labels were based on expert subjective assessment, which may not capture all aspects of image interpretability. Similarly, manual segmentations in all chapters are subject to inter- and intra-observer variability. While our quality-aware model exceeded expert agreement in some cases, the reliance on a single consensus annotation as *truth* is an inherent limitation of

supervised learning in medicine. Future work, as discussed in Section 6.3.1, should aim to mitigate this by developing self-supervised and physics-informed methods for generating objective quality metrics.

Beyond Technical Performance: This thesis focuses on technical and methodological contributions. However, technical performance alone is insufficient for clinical adoption. Substantial non-technical barriers remain, including navigating complex regulatory approval pathways (e.g., FDA, CE marking), addressing issues of liability and ethics, and ensuring seamless and secure integration with existing Picture Archiving and Communication Systems (PACS) and hospital information systems. These crucial aspects of clinical translation were outside the scope of this work. However, future development should incorporate human-in-the-loop and explainability frameworks, as detailed in Section 6.3.2, to help bridge this gap and facilitate clinical adoption.

6.3 Future Research Directions

The foundations established by this thesis open several critical research pathways. Future work should systematically address the limitations identified above while building upon the specific methodological advances demonstrated here.

6.3.1 Immediate Extensions: Strengthening Core Methodologies

The frameworks developed in this thesis provide a strong foundation for several immediate research extensions aimed at enhancing their analytical depth, robustness, and deployability.

First, the scope of analysis can be expanded from static anatomical mapping to dynamic and quantitative assessment. The state-of-the-art 3D vessel wall segmentation model (Chapter 4) provides the anatomical foundation for developing automated modules that extract clinically vital biomarkers, such as total plaque volume, wall thickness, and luminal stenosis. This moves the output from a simple mask to actionable clinical data. This concept can be further extended in two dimensions: longitudinally, by analyzing serial MRI scans to precisely track disease progression or therapeutic response over time, and temporally, by adapting the frameworks to 4D data like ultrasound videos or dynamic MRI to capture physiological processes such as vessel wall motion and colon peristalsis.

Second, a major research thrust should focus on generalizing and automating the

principle of quality-aware robustness. The framework introduced in Chapter 5) for 2D ultrasound is a powerful but specific solution. A key next step is to extend this concept into a unified, cross-modality architecture that is robust for both 2D ultrasound and 3D MRI. This would require defining modality-specific quality metrics, such as motion artifacts in MRI. To overcome the current reliance on subjective expert labels, future work should pivot toward physics-informed and self-supervised learning. This would enable models to predict objective, physics-based quality metrics directly from the image data, creating a fully automated and more scalable pipeline for robust analysis.

Finally, practical deployment requires continued optimization for diverse clinical environments. Building on the efficiency gains achieved in Chapter 2, these frameworks must be tailored for real-world hardware constraints. This involves a dual strategy: aggressive model compression and quantization for deployment on computationally limited hardware like portable ultrasound devices (edge deployment), alongside ensuring rapid and scalable throughput for processing large 3D MRI cohorts on hospital-grade servers (cloud or HPC deployment).

6.3.2 Addressing Key Limitations for Clinical Translation

Successfully translating the frameworks developed in this thesis into routine clinical tools requires addressing two fundamental challenges: proving their generalizability across diverse real-world settings and ensuring they are designed to be trustworthy and integrable into clinical workflows.

First and foremost, the highest priority is to move beyond the single-institution datasets that formed the basis of this research. Rigorous, prospective validation studies must be conducted in collaboration with multiple international clinical sites. This is the only way to test the models' performance on diverse patient populations and a wide array of imaging equipment. Such an endeavor will inevitably require the development of robust domain adaptation and federated learning techniques, which will allow the models to adapt to new clinical environments while preserving patient privacy [18].

Alongside this large-scale validation, building clinical trust and facilitating regulatory approval is equally critical. This necessitates designing future systems not as autonomous "black boxes", but as collaborative, human-in-the-loop tools with comprehensive explainability frameworks. While segmentation outputs inherently offer a degree of spatial explainability by visualizing the identified anatomical structures, this transparency must be enhanced to meet the demands of clinical decision-making. This involves creating sophisticated user interfaces that allow clinicians to interactively

review, correct, and validate AI-generated masks. These systems should be augmented with real-time model confidence scores and pixel-level uncertainty quantification, using visual cues to highlight regions of low confidence. This approach not only directs expert attention to where it is needed most but also enables seamless correction workflows that can improve model performance over time through active learning paradigms.

6.4 Concluding Perspective

This thesis has demonstrated that the fundamental challenges preventing the clinical deployment of medical imaging AI —efficiency, clinical integration, precision, and robustness to data variability— can be systematically addressed through principled methodological innovation. The progression from an efficient end-to-end framework, through clinical risk integration, to high-precision 3D segmentation and quality-aware robustness provides a comprehensive roadmap for developing clinically translatable AI.

The future of diagnostic imaging lies not in replacing human expertise, but in augmenting it with AI systems that are robust, interpretable, and seamlessly integrated into clinical workflows. By addressing the core challenges of real-world deployment across different modalities, the frameworks developed here contribute to a future where advanced diagnostic capabilities are accessible globally, empowering clinicians and improving patient outcomes. The path from laboratory success to global health impact requires continued, rigorous collaboration between computer scientists, clinicians, engineers, and regulatory bodies. This thesis provides validated building blocks for that journey, but the ultimate measure of success will be improved patient outcomes in diverse clinical settings worldwide.

Appendix A

Contributions

Beyond the core research presented in this thesis, my doctoral work has involved significant contributions to the scientific community through articles in high-impact journals, presentations at major international conferences, and participation in competitive medical imaging challenges. These efforts, detailed below, highlight my engagement with the broader research landscape and my commitment to advancing the field of medical image analysis.

A.1 Publications in Indexed Journals

1. **An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery.**
   Journal: Computer Methods and Programs in Biomedicine, 223, 106954, 2022.
   DOI: `https://doi.org/10.1016/j.cmpb.2022.106954`
   Impact factor: 6.1 (Q1 in Computer Science, Theory & Methods.)
   Authors: Gago, L., del Mar Vila, M., Grau, M., Remeseiro, B., & Igual, L.

2. **Deep-stratification of the cardiovascular risk by ultrasound carotid artery images.**
   Journal: Biomedical Signal Processing and Control, 91, 106035, 2024.
   DOI: `https://doi.org/10.1016/j.bspc.2024.106035`
   Impact factor: 4.9 (Q2 in Biomedical Engineering)
   Authors: Vila, M. del M., Gago, L., Pérez-Sánchez, P., Grau, M., Remeseiro, B., & Igual, L.

3. **Bridging the Quality Gap: Robust Colon Wall Segmentation in Noisy Transabdominal Ultrasound.**
   Journal: Computers in Biology and Medicine, Volume 197, 111077, 2025.
   DOI: `https://doi.org/10.1016/j.compbiomed.2025.111077`
   Impact factor: 6.3 (Q1 in Computer Science, Interdisciplinary Applications)
   Authors: Gago, L., Fernández González, M. Á., Engelmann, J., Remeseiro, B., & Igual, L.

4. **Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation (CURVAS) challenge results**
   Journal: Computers In Biology And Medicine, 197B, 2025.
   DOI: `https://doi.org/10.1016/j.compbiomed.2025.111024`
   Impact factor: 6.3 (Q1 in Computer Science, Interdisciplinary Applications)
   Authors: Riera-Marin, M., Rodriguez-Comas, J., May, M. S., Pan, Z., Zhou, X., Liang, X., Erick, F. X., Prenner, A., Hemon, C., Boussot, V., Dillenseger, J.-L., Nunes, J.-C., Qayyum, A., Mazher, M., Niederer, S. A., Kushibar, K., Martin-Isla, C., Radeva, P., Lekadir, K., Barfoot, T., Herrera, L. C., Glocker, B., Vercauteren, T., Gago, L., Engelmann, J., Kleiss, J.-M., Aubanell, A., Antolin, A., Garcia-Lopez, J., Gonzalez Ballester, M. A., & Galdran, A.

5. **Applicability of oculomics for individual risk prediction: Repeatability and robustness of retinal Fractal Dimension using DART and AutoMorph.**
   Journal: Investigative Ophthalmology & Visual Science, 65, 10, 2025.
   DOI: `https://doi.org/10.1167/iovs.65.6.10`
   Impact factor: 5.0 (Q1 in Ophthalmology)
   Authors: Engelmann, J., Moukaddem, D., Gago, L., Strang, N., & Bernabeu, M. O.

6. **Percutaneous Ultrasound-Guided Radiofrequency Ablation as a Therapeutic Approach for the Management of Insulinomas and Associated Metastases in Dogs.**
   Journal: Animals, 14(22), 3301, 2024.
   DOI: `https://doi.org/10.3390/ani14223301`
   Impact factor: 2.7 (Q1 in Veterinary Sciences)
   Authors: Alférez, M. D., Corda, A., Blas, I. de, Gago, L., Fernandes, T., Rodríguez-Piza, I., Balañá, B., Corda, F., & Gómez Ochoa, P.

7. **Computed Tomography-Guided Radiofrequency Ablation of Nasal Carcinomas in Dogs.**
   Journal: Animals, 14(24), 3682, 2024.
   DOI: `https://doi.org/10.3390/ani14243682`
   Impact factor: 2.7 (Q1 in Veterinary Sciences)
   Authors: Alférez, M. D., Corda, A., Blas, I. de, Gago, L., Fernandes, T., Rodríguez-Piza, I., Balañá, B., Pentcheva, P., Caruncho, J., Barbero-Fernández, A., Llinás, J., Rivas, D., Escudero, A., & Gómez-Ochoa, P.

Under peer-review:

1. **Context-Aware Multilevel EfficientNet-UNet++ for Precise 3-D Carotid Vessel-Wall Segmentation**
   Submitted to: Expert Systems With Applications
   Impact factor: 7.5 (Q1 in Computer Science, Artificial Intelligence)
   Authors: Gago, L., Remeseiro, B., & Igual, L.

A.2 Other Publications

1. **Self-consistent deep approximation of retinal traits for robust and highly efficient vascular phenotyping of retinal colour fundus images.**
   Reference: Communications in Computer and Information Science, vol. 2240. Springer, Cham, 2024.
   DOI: `https://doi.org/10.1007/978-3-031-79103-1_22`
   Authors: Gago, L., Remeseiro, B., Igual, L., Storkey, A., Bernabeu, M. O., & Engelmann, J.

2. **An Ultra-efficient Method for Real-Time Ultra-widefield Fundus Image Quality Assessment**
   Reference: Lecture Notes in Computer Science, vol. 15597. Springer, Cham, 2025.
   DOI: `https://doi.org/10.1007/978-3-031-89388-9_11`
   Authors: Engelmann, J. & Gago, L.

3. **Ultra-fast Detection of Referable Diabetic Retinopathy and Macular Edema in Ultra-widefield Fundus Imaging Using a Unified Risk Score**
   Reference: Lecture Notes in Computer Science, vol. 15597. Springer, Cham, 2025.

DOI: https://doi.org/10.1007/978-3-031-89388-9_12
Authors: Engelmann, J. & Gago, L.

4. **Repeatability and measurement noise of retinal Fractal Dimension using Deep Approximation of Retinal Traits (DART) and their relationship with image quality**
   Reference: Investigative Ophthalmology & Visual Science, vol. 65, 5964, 2024.
   Link: https://iovs.arvojournals.org/article.aspx?articleid=2795671
   Authors: Engelmann, J., Gago, L., Moukaddem, D., Burke, J., Storkey, A., Bernabeu, M., & Strang, N. C.

5. **Efficient ocular toxoplasmosis detection with RETFound-Green, a retinal foundation model**
   Reference: Investigative Ophthalmology & Visual Science, vol. 66, 4199, 2025.
   Link: https://iovs.arvojournals.org/article.aspx?articleid=2805037
   Authors: Gago, L., Engelmann, J., Sevgi, M., & Elbatel, M.

6. **Developing a high-performance, smartphone-ready glaucoma AI model with RETFound-Green**
   Reference: Investigative Ophthalmology & Visual Science, vol. 66, 2717, 2025.
   Link: https://iovs.arvojournals.org/article.aspx?articleid=2805096
   Authors: Kitema, G. F., Engelmann, J., Morley, M., Sevgi, M., Ong, A. Y., Ruffell, E., Merle, D. A., Gago, L., Wagner, S., & Keane, P. A.

7. **RETFound-Mobile: Running a high-performance retinal foundation model on a regular smartphone**
   Reference: Investigative Ophthalmology & Visual Science, vol. 66, 4636, 2025.
   Link: https://iovs.arvojournals.org/article.aspx?articleid=2804797
   Authors: Engelmann, J., Kitema, G. F., Morley, M., Sevgi, M., Ong, A. Y., Ruffell, E., Merle, D. A., Gago, L., Wagner, S., & Keane, P. A.

Accepted but not yet published:

1. **Input Simplification Impact on Robustness for Targeted Therapy Subtypes in Breast MRI Segmentation AI**
   Reference: Lecture Notes in Computer Science
   Authors: Rodriguez, A. B., Remeseiro, B., Engelmann, J., & Gago, L.

A.3 Conference Presentations

1. **Input Simplification Impact on Robustness for Targeted Therapy Subtypes in Breast MRI Segmentation AI**
   Conference: Deep-Breath 2025: 2nd Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, MICCAI 2025 Workshop
   Date: September 23, 2025

2. **Self-consistent deep approximation of retinal traits for robust and highly efficient vascular phenotyping of retinal colour fundus images**
   Conference: 2024 Data Science and AI Symposium, Harvard Ophthalmology and Mass Eye and Ear
   Date: September 27, 2024

3. **Visual Field test prediction with multi-output-regularised RandomForest without optical coherence tomography**
   Conference: Structural-Functional Transition in Glaucoma Assessment 2, MICCAI 2024 Challenge Workshop
   Date: October 10, 2024

4. **An Ultra-efficient Method for Real-Time Ultra-widefield Fundus Image Quality Assessment**
   Conference: Ultra widefield fundus imaging for diabetic retinopathy 2024, MICCAI 2024 Challenge Workshop
   Date: October 6, 2024

5. **Self-consistent deep approximation of retinal traits for robust and highly efficient vascular phenotyping of retinal colour fundus images**
   Conference: MICCAI Student Board Emerge Workshop, MICCAI 2024
   Date: October 6, 2024

6. **SFOOD (Simple Features OOD)**
   Conference: MSB Emerge Workshop, MICCAI 2023
   Date: October 12, 2023

7. **Enhancing 3D Carotid Artery MRI Segmentation: Multilevel Efficientnet-Unet++ with Contextual Information**
   Conference: COSMOS 2022 Challenge Workshop, MICCAI 2022
   Date: September 18, 2022

A.4 Awards and Honors

1. **1st Place:** Achieved first place in the MICCAI 2024 challenge on *Structural-Functional Transition in Glaucoma Assessment*.
2. **1st Place:** Achieved first place in the MICCAI 2024 challenge on *Structural-Functional Transition in Glaucoma Assessment* for Sensitivity Map Prediction.
3. **1st Place:** Achieved first place in the MICCAI 2024 challenge on *Structural-Functional Transition in Glaucoma Assessment* for Prediction of Mean Deviation.
4. **3rd Place:** Achieved third place in the MICCAI 2024 challenge on *Structural-Functional Transition in Glaucoma Assessment* for Pattern Deviation Probability Map Prediction.
5. **Outstanding Contribution Award:** Received the outstanding contribution award in the MICCAI 2024 challenge on *Ultra widefield fundus imaging for diabetic retinopathy*.
6. **2nd Place:** Achieved second place in the MICCAI 2024 challenge on *Ultra widefield fundus imaging for diabetic retinopathy* for Image Quality Assessment of Ultra-widefield Fundus.
7. **2nd Place:** Achieved second place in the MICCAI 2024 challenge on *Ultra widefield fundus imaging for diabetic retinopathy* for Identification of Diabetic Macular Edema.
8. **2nd Place:** Achieved second place in the MICCAI MSB Emerge Workshop 2024.
9. **Top-5 Finalist:** Achieved top 5 finalist position in the COSMOS Grand Challenge, MICCAI 2022.

Others

1. **Full Travel Grant:** 2024 Data Science and AI Symposium, Harvard Ophthalmology and Mass Eye and Ear.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[2] Eric J Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature medicine* 25.1 (2019), pp. 44–56.

[3] Stephen F. Weng et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" In: *PLOS ONE* 12.4 (2017), pp. 1–14.

[4] Bharath Ambale-Venkatesh et al. "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis". In: *Circulation Research* 121.9 (2017), pp. 1092–1101.

[5] Zeynettin Akkus et al. "A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow". In: *Journal of the American College of Radiology* 16.9 (2019), pp. 1318–1328.

[6] Christian Maaser et al. "Intestinal ultrasound for monitoring therapeutic response in patients with ulcerative colitis: results from the TRUST&UC study". In: *Gut* 69.9 (2020), pp. 1629–1636.

[7] Peter Libby. "The changing landscape of atherosclerosis". In: *Nature* 592.7855 (2021), pp. 524–533.

[8] P.-J. Touboul et al. "Mannheim carotid intima-media thickness and plaque consensus (2004-2006-2011)." In: *Cardiovascular Diseases* 34.4 (2012), pp. 290–296.

[9] Daniel H. O'Leary et al. "Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults". In: *New England Journal of Medicine* 340.1 (1999), pp. 14–22.

[10] Christos P. Loizou. "A review of ultrasound common carotid artery image and video segmentation techniques". In: *Medical and Biological Engineering and Computing* 52.12 (Nov. 2014), pp. 1073–1093.

[11] Lillian Du and Christina Ha. "Epidemiology and pathogenesis of ulcerative colitis". In: *Gastroenterology Clinics* 49.4 (2020), pp. 643–654.

[12] Ramona Leenings et al. "C-TRUS: A Novel Dataset and Initial Benchmark for Colon Wall Segmentation in Transabdominal Ultrasound". In: *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer. 2024, pp. 101–111.

[13] Maria Grau et al. "Carotid Intima-media Thickness in the Spanish Population: Reference Ranges and Association With Cardiovascular Risk Factors". In: *Revista Española de Cardiologia* 65.12 (2012), pp. 1086–1093.

[14] Huijun Chen et al. *Carotid Vessel Wall Segmentation and Atherosclerosis Diagnosis Challenge*. 2022. URL: `https://vessel-wall-segmentation-2022.grand-challenge.org/`.

[15] Phillip Gu et al. "Radiomics-based analysis of intestinal ultrasound images for inflammatory bowel disease: a feasibility study". In: *Crohn's & Colitis 360* 6.2 (2024), otae034.

[16] Joseph W Goodman. "Some fundamental properties of speckle". In: *JOSA* 66.11 (1976), pp. 1145–1150.

[17] Muhammad Moinuddin et al. "Medical ultrasound image speckle reduction and resolution enhancement using texture compensated multi-resolution convolution neural network". In: *Frontiers in Physiology* 13 (2022), p. 961571.

[18] Risheng Wang et al. "Medical image segmentation using deep learning: A survey". In: *IET image processing* 16.5 (2022), pp. 1243–1267.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. 2015, pp. 234–241.

[20] Aldons J. Lusis. "Atherosclerosis". In: *Nature* 407 (2000), pp. 233–241.

[21] Young Mi Hong. "Atherosclerotic cardiovascular disease beginning in childhood". In: *Korean Circulation Journal* 40.1 (2010), pp. 1–9.

[22] Trajen Head, Sylvia Daunert, and Pascal J Goldschmidt-Clermont. "The aging risk and atherosclerosis: a fresh look at arterial homeostasis". In: *Frontiers in Genetics* 8 (2017), p. 216.

[23] Thomas A. Gaziano et al. "Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries". In: *Current Problems in Cardiology* 35.2 (2010), pp. 72–115.

[24] Adelina Doltra et al. "Magnetic resonance imaging of cardiovascular fibrosis and inflammation: from clinical practice to animal studies and back". In: *BioMed Research International* 2013.10 (2013), pp. 1–2.

[25] Michiel L. Bots et al. "Common carotid intima-media thickness and risk of stroke and myocardial infarction: The Rotterdam Study". In: *Circulation* 96.5 (1997), pp. 1432–1437.

[26] C. P. Loizou et al. "M-mode state based identification in ultrasound videos of the atherosclerotic carotid plaque". In: *4th International Symposium on Communications, Control, and Signal Processing*. 2010, pp. 1–6.

[27] Maria del Mar Vila et al. "Semantic segmentation with DenseNets for carotid artery ultrasound plaque segmentation and CIMT estimation". In: *Artificial Intelligence in Medicine* 103 (2020), p. 101784.

[28] Filippo Molinari et al. "Completely automated multiresolution edge snapper—a new technique for an accurate carotid ultrasound IMT measurement: clinical validation and benchmarking on a multi-institutional database". In: *IEEE Transactions on Image Processing* 21.3 (2011), pp. 1211–1222.

[29] M. C. Bastida-Jumilla et al. "Frequency-domain active contours solution to evaluate intima-media thickness of the common carotid artery". In: *Biomedical Signal Processing and Control* 16 (2015), pp. 68–79.

[30] C. P. Loizou et al. "Snakes based segmentation of the common carotid artery intima media". In: *Medical & Biological Engineering & Computing* 45.1 (2007), pp. 35–49.

[31] Chunjun Qian and Xiaoping Yang. "An integrated method for atherosclerotic carotid plaque segmentation in ultrasound image". In: *Computer Methods and Programs in Biomedicine* 153 (2018), pp. 19–32.

[32] Rosa-María Menchón-Lara and José-Luis Sancho-Gómez. "Fully automatic segmentation of ultrasound common carotid artery images based on machine learning". In: *Neurocomputing* 151 (2015), pp. 161–167.

[33] Mainak Biswas et al. “Deep learning strategy for accurate carotid intima-media thickness measurement: an ultrasound study on Japanese diabetic cohort”. In: *Computers in Biology and Medicine* 98 (2018), pp. 100–117.

[34] Mainak Biswas et al. “Two-stage artificial intelligence model for jointly measurement of atherosclerotic wall thickness and plaque burden in carotid ultrasound: A screening tool for cardiovascular/stroke risk assessment”. In: *Computers in Biology and Medicine* 123 (2020), p. 103847.

[35] Mainak Biswas et al. “A review on joint carotid intima-media thickness and plaque area measurement in ultrasound for cardiovascular/stroke risk monitoring: Artificial Intelligence framework”. In: *Journal of Digital Imaging* 34.3 (2021), pp. 581–604.

[36] Nobutaka Ikeda et al. “Automated segmental-IMT measurement in thin/thick plaque with bulb presence in carotid ultrasound from multiple scanners: stroke risk assessment”. In: *Computer Methods and Programs in Biomedicine* 141 (2017), pp. 73–81.

[37] Sheng Lian et al. “APRIL: Anatomical prior-guided reinforcement learning for accurate carotid lumen diameter and intima-media thickness measurement”. In: *Medical Image Analysis* 71 (2021), p. 102040.

[38] Elisabetta Bianchini et al. “Functional and Structural Alterations of Large Arteries: Methodological Issues”. In: *Current Pharmaceutical Design* 19.13 (2013), pp. 2390–2400.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. 2015, pp. 234–241.

[40] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *32nd International Conference on Machine Learning*. Vol. 37. 2015, pp. 448–456.

[41] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958.

[42] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[43] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.

[44] Christos P. Loizou et al. "Comparative evaluation of despeckle filtering in ultrasound imaging of the carotid artery". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 52.10 (2005), pp. 1653–1669.

[45] Tsung Yi Lin et al. "Focal Loss for Dense Object Detection". In: *IEEE International Conference on Computer Vision*. 2017, pp. 2999–3007.

[46] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2951–2959.

[47] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *27th International Conference on Machine Learning*. 2010, pp. 807–814.

[48] Abadi Martín et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016), pp. 1–19.

[49] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations*. 2015, pp. 1–15.

[50] Jaume Marrugat et al. "Validez relativa de la estimación del riesgo cardiovascular a 10 años en una cohorte poblacional del estudio REGICOR". In: *Revista Española de Cardiología* 64.15 (2011), pp. 385–394.

[51] Sr D'Agostino Ralph B. et al. "Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation". In: *JAMA* 286.2 (2001), pp. 180–187.

[52] Jaume Marrugat et al. "Derivation and validation of a set of 10-year cardiovascular risk predictive functions in Spain: The FRESCO Study". In: *Preventive Medicine* 61 (2014), pp. 66–74. ISSN: 0091-7435.

[53] R.M. Conroy et al. "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project". In: *European Heart Journal* 24.11 (June 2003), pp. 987–1003. ISSN: 0195-668X.

[54] Julia Hippisley-Cox et al. "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2". In: *BMJ* 336.1475 (2006).

[55] M Grau and J Marrugat. "Funciones de riesgo en la prevención primaria de las enfermedades cardiovasculares". In: *Revista Española de Cardiología* 61.4 (2008), pp. 404–416.

[56] J Marrugat et al. "Comparing the Predictive Powers of Survival Models Using Harrell's C or Somers' D". In: *J Epidemiol Community Health* 57.8 (2003), pp. 634–8.

[57] Luca Saba et al. "Plaque tissue morphology-based stroke risk stratification using carotid ultrasound: a polling-based PCA learning paradigm". In: *Journal of Medical Systems* 41.6 (2017), p. 98.

[58] Tadashi Araki et al. "Stroke risk stratification and its validation using ultrasonic echolucent carotid wall plaque morphology: a machine learning paradigm". In: *Computers in Biology and Medicine* 80 (2017), pp. 77–96.

[59] Marie-Louise M. Gronholdt et al. "Ultrasonic Echolucent Carotid Plaques Predict Future Strokes". In: *Circulation* 104.1 (2001), pp. 68–73.

[60] Yoko Irie et al. "The utility of ultrasonic tissue characterization of carotid plaque in the prediction of cardiovascular events in diabetic patients". In: *Atherosclerosis* 230.2 (2013), pp. 399–405.

[61] Narendra N Khanna et al. "Performance evaluation of 10-year ultrasound image-based stroke/cardiovascular (CV) risk calculator by comparing against ten conventional CV risk calculators: a diabetic study". In: *Computers in Biology and Medicine* 105 (2019), pp. 125–143.

[62] C Mitchell et al. "Ultrasound carotid plaque features, cardiovascular disease risk factors and events: The Multi-Ethnic Study of Atherosclerosis". In: *Atherosclerosis* 276 (2018), pp. 195–202.

[63] Efthyvoulos C Kyriacou et al. "Prediction of high-risk asymptomatic carotid plaques based on ultrasonic image features". In: *IEEE Transactions on Information Technology in Biomedicine* 16.5 (2012), pp. 966–973.

[64] R Nakanishi et al. "Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths". In: *JACC Cardiovasc Imaging* 14.3 (2021), pp. 615–625.

[65] BK Tamarappoo et al. "Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: A prospective study". In: *Atherosclerosis* (2021), pp. 76–82.

[66] D Gonzalo-Calvo et al. "Soluble low-density lipoprotein receptor-related protein 1 as a biomarker of coronary risk: Predictive capacity and association with clinical events". In: *Atherosclerosis* 287 (2019), pp. 93–99.

[67] S Winther et al. "Coronary Calcium Scoring Improves Risk Prediction in Patients With Suspected Obstructive Coronary Artery Disease". In: *Journal of the American College of Cardiology* 80.21 (2022), pp. 1965–1977.

[68] Gangireddy Narendra Kumar Reddy, M. Sabarimalai Manikandan, and Ram Bilas Pachori. "Automated Hilbert Envelope Based Respiration Rate Measurement from PPG Signal for Wearable Vital Signs Monitoring Devices". In: *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*. 2022, pp. 1–6. DOI: 10.1109/ICAIoT57170.2022.10121855.

[69] Ashish Sharma et al. "Accurate tunable-Q wavelet transform based method for QRS complex detection". In: *Computers & Electrical Engineering* 75 (2019), pp. 101–111. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2019.01.025. URL: https://www.sciencedirect.com/science/article/pii/S0045790618322316.

[70] Lucas Gago et al. "An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery". In: *Computer Methods and Programs in Biomedicine* 223 (2022), p. 106954.

[71] N. E. Breslow. "Analysis of Survival Data under the Proportional Hazards Model". In: *International Statistical Review / Revue Internationale de Statistique* 43.1 (1975), pp. 45–57. ISSN: 03067734, 17515823. (Visited on 01/27/2023).

[72] Frank E. Harrell, K L Lee, and Daniel B. Mark. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." In: *Statistics in medicine* 15 4 (1996), pp. 361–87.

[73] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (2010), pp. 433–459.

[74] Bendix Carstensen et al. *Epi: A Package for Statistical Analysis in Epidemiology*. R package version 2.44. 2021. URL: https://CRAN.R-project.org/package=Epi.

[75] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[76] Eric de Groot et al. "Measurement of arterial wall thickness as a surrogate marker for atherosclerosis". In: *Circulation* 109.23 suppl 1 (2004), pp. III–33.

[77] Jr Harrell Frank E. et al. "Evaluating the Yield of Medical Tests". In: *JAMA* 247.18 (May 1982), pp. 2543–2546. ISSN: 0098-7484.

[78] Michael J. Pencina, Ralph B. D'Agostino, and Ramachandran S. Vasan. "Statistical methods for assessment of added usefulness of new biomarkers". In: *Clinical Chemistry and Laboratory Medicine* 48.12 (2010), pp. 1703–1711.

[79] Eisuke Inoue. *nricens: NRI for Risk Prediction Models with Time to Event and Binary Response Data*. R package version 1.6. 2018. URL: `https://CRAN.R-project.org/package=nricens`.

[80] Gareth James et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.

[81] Huilin Zhao et al. "Assessment of carotid artery atherosclerotic disease by using three-dimensional fast black-blood MR imaging: comparison with DSA". In: *Radiology* 274.2 (2015), pp. 508–516.

[82] Markus Henningsson et al. "Black-Blood Contrast in Cardiovascular MRI". In: *Journal of Magnetic Resonance Imaging* 55.1 (2022), pp. 61–80.

[83] Baocheng Chu et al. "Reproducibility of carotid atherosclerotic lesion type characterization using high resolution multicontrast weighted cardiovascular magnetic resonance". In: *Journal of Cardiovascular Magnetic Resonance* 8.6 (2006), pp. 793–799.

[84] Chun Yuan et al. "Closed contour edge detection of blood vessel lumen and outer wall boundaries in black-blood MR images". In: *Magnetic Resonance Imaging* 17.2 (1999), pp. 257–266.

[85] Gareth Adams et al. "Algorithm for quantifying advanced carotid artery atherosclerosis in humans using MRI and active contours". In: *Medical Imaging 2002: Image Processing*. Vol. 4684. SPIE. 2002, pp. 1448–1457.

[86] Hunter R Underhill et al. "Automated measurement of mean wall thickness in the common carotid artery by MRI: a comparison to intima-media thickness by B-mode ultrasound". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 24.2 (2006), pp. 379–387.

[87] Andres M Arias Lorza et al. "Maximization of regional probabilities using Optimal Surface Graphs: Application to carotid artery segmentation in MRI". In: *Medical Physics* 45.3 (2018), pp. 1159–1169.

[88] Shan Gao et al. "Quantification of common carotid artery and descending aorta vessel wall thickness from MR vessel wall imaging using a fully automated processing pipeline". In: *Journal of Magnetic Resonance Imaging* 45.1 (2017), pp. 215–228.

[89] Li Chen et al. "Automatic segmentation of carotid vessel wall using convolutional neural network". In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine* 96 (2018), pp. 2017–2018.

[90] Li Chen et al. "Automated artery localization and vessel wall segmentation using tracklet refinement and polar conversion". In: *IEEE Access* 8 (2020), pp. 217603–217614.

[91] Dieuwertje Alblas, Christoph Brune, and Jelmer M Wolterink. "Deep-learning-based carotid artery vessel wall segmentation in black-blood MRI using anatomical priors". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 237–244.

[92] COSMOS 2022. *Training dataset for Carotid Vessel Wall Segmentation and Atherosclerosis Diagnosis Challenge, MICCAI 2022*. 2022. DOI: 10.5281/zenodo.6481870. URL: https://doi.org/10.5281/zenodo.6481870.

[93] COSMOS 2022. *Testing dataset for Carotid Vessel Wall Segmentation and Atherosclerosis Diagnosis Challenge, MICCAI 2022*. 2022. DOI: 10.5281/zenodo.6843257. URL: https://doi.org/10.5281/zenodo.6843257.

[94] Shishuai Hu, Zehui Liao, and Yong Xia. "Label Propagation for 3D Carotid Vessel Wall Segmentation and Atherosclerosis Diagnosis". In: *arXiv preprint arXiv:2208.13337* (2022).

[95] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211.

[96] Haoxuan Li et al. "DBF-UNet: A Two-Stage Framework for Carotid Artery Segmentation with Pseudo-Label Generation". In: *arXiv preprint arXiv:2504.00908* (2025).

[97] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026.

[98] Junlong Cheng et al. *SAM-Med2D*. 2023. arXiv: `2308.16184 [cs.CV]`.

[99] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66.

[100] Zongwei Zhou et al. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[101] Abien Fred Agarap. "Deep Learning using Rectified Linear Units (ReLU)". In: *arXiv preprint arXiv:1803.08375* (2018).

[102] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241.

[103] William Kerwin et al. "Magnetic resonance imaging of carotid atherosclerosis: plaque analysis". In: *Topics in Magnetic Resonance Imaging* 18.5 (2007), pp. 371–378.

[104] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems* 32 (2019).

[105] Tobias Saam et al. "Quantitative evaluation of carotid plaque composition by in vivo MRI". In: *Arteriosclerosis, thrombosis, and vascular biology* 25.1 (2005), pp. 234–239.

[106] Matthias W Lorenz et al. "Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis". In: *Circulation* 115.4 (2007), pp. 459–467.

[107] Bruce A Wasserman et al. "MRI measurements of carotid plaque in the atherosclerosis risk in communities (ARIC) study: methods, reliability and descriptive statistics". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 31.2 (2010), pp. 406–415.

[108] Steven Bots et al. "Intestinal ultrasound to assess disease activity in ulcerative colitis: development of a novel UC-ultrasound index". In: *Journal of Crohn's and Colitis* 15.8 (2021), pp. 1264–1271.

[109] C Pahl et al. "Performance observation of gabor filter for wall thickness measurement of human colon based on ultrasound image". In: *International Journal of Information and Electronics Engineering* 4.2 (2014), p. 171.

[110] NORAYATI Nordin et al. "Wall thickness measurement of colon based on ultrasound image segmentation". In: *1st WSEAS International Conference on Biomedicine and Health Engineering*. 2012, pp. 324–329.

[111] Shi Yin et al. "Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks". In: *Medical Image Analysis* 60 (2020), p. 101602.

[112] Mohammed Yusuf Ansari et al. "Dense-PSP-UNet: a neural network for fast inference liver ultrasound segmentation". In: *Computers in Biology and Medicine* 153 (2023), p. 106478.

[113] Zeynettin Akkus et al. "Fully automated segmentation of bladder sac and measurement of detrusor wall thickness from transabdominal ultrasound images". In: *Sensors* 20.15 (2020), p. 4175.

[114] Tao Peng et al. "H-ProMed: Ultrasound image segmentation based on the evolutionary neural network and an improved principal curve". In: *Pattern Recognition* 131 (2022), p. 108890.

[115] Hariharan Ravishankar et al. "Learning and incorporating shape models for semantic segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 203–211.

[116] Danilo Avola et al. "Ultrasound medical imaging techniques: a survey". In: *ACM Computing Surveys* 54.3 (2021), pp. 1–38.

[117] Yu Wang et al. "Deep learning in medical ultrasound image analysis: a review". In: *IEEE Access* 9 (2021), pp. 54310–54324.

[118] Reza Azad et al. "Contextual attention network: Transformer meets u-net". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2022, pp. 377–386.

[119] Moein Heidari et al. "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation". In: *IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6202–6212.

[120] Gongping Chen et al. "AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images". In: *IEEE Transactions on Medical Imaging* 42.5 (2022), pp. 1289–1300.

[121] Gongping Chen et al. "Rethinking the unpretentious U-net for medical ultrasound image segmentation". In: *Pattern Recognition* 142 (2023), p. 109728.

[122] Gongping Chen, Yu Dai, and Jianxun Zhang. "C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation". In: *Computer Methods and Programs in Biomedicine* 225 (2022), p. 107086.

[123] Yating Ling et al. "Mtanet: Multi-task attention network for automatic medical image segmentation and classification". In: *IEEE Transactions on Medical Imaging* 43.2 (2023), pp. 674–685.

[124] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[125] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[126] Meng-Hao Guo et al. "Segnext: Rethinking convolutional attention design for semantic segmentation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1140–1156.

[127] Kaiming He et al. "Mask R-CNN". In: *IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.

[128] Michael Yeung et al. "Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation". In: *Journal of Digital Imaging* 36.2 (2023), pp. 739–752.

[129] Zongwei Zhou et al. "Models genesis: Generic autodidactic models for 3d medical image analysis". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 384–393.

[130] Xu Chen et al. "Learning active contour models for medical image segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 11632–11640.

[131] Maria Tirindelli et al. "Rethinking ultrasound augmentation: A physics-inspired approach". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer. 2021, pp. 690–700.

[132] H Zhang et al. "mixup: Beyond empirical risk management". In: *6th Int. Conf. Learning Representations (ICLR)*. 2018, pp. 1–13.

[133] Zhuang Liu et al. "A convnet for the 2020s". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

[134] Ma Yi-de, Liu Qing, and Qian Zhi-Bai. "Automated image segmentation using improved PCNN model based on cross-entropy". In: *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE. 2004, pp. 743–746.

[135] Carole H Sudre et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer. 2017, pp. 240–248.

[136] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *International Conference on Machine Learning*. Pmlr. 2013, pp. 1310–1318.

[137] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization. 7th Int". In: *Conf. Learn. Represent. ICLR*. 2019.

[138] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 ieee information theory workshop (itw)*. Ieee. 2015, pp. 1–5.

[139] Saikat Roy et al. "Mednext: transformer-driven scaling of convnets for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 405–415.

[140] Zhuo Zhang et al. "CI-UNet: melding convnext and cross-dimensional attention for robust medical image segmentation". In: *Biomedical engineering letters* 14.2 (2024), pp. 341–353.

[141] Alex Kendall, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[142] Soon Hoe Lim et al. "Noisy feature mixup". In: *Proceedings of the 10th International Conference on Learning Representations*. 2022.

[143] Lina Pedraza et al. “An open access thyroid ultrasound image database”. In: *10th International symposium on medical information processing and analysis*. Vol. 9287. SPIE. 2015, pp. 188–193.