Doctoral Dissertation
Doctoral Program in Bioengineering and Medical-Surgical Sciences (36th Cycle)

Deep Learning in B-mode Ultrasound Imaging: from Discriminative Analysis to Generative Models

By

Francesco Marzola

******

Supervisors:
Prof. Kristen M. Meiburger
Prof. Massimo Salvi
Prof. Filippo Molinari

Doctoral Examination Committee:
Prof. Spyretta Golemati, National and Kapodistrian University of Athens
Prof. Ruud van Sloun, Eindhoven University of Technology

Politecnico di Torino
2024

Declaration

I hereby declare that the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

This thesis is licensed under a Creative Commons License, Attribution - Noncommercial - NoDerivative Works 4.0 International: see www.creativecommons.org. The text may be reproduced for non-commercial purposes, provided that credit is given to the original author.

Turin, 19/04/2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

AI use Disclosure

In the writing process of this thesis the following tools have been used to improve readability and language:

- Grammarly (Grammarly, Inc.)
- ChatGPT (OpenAI)

Turin, 19/04/2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Summary

B-mode ultrasound imaging stands as a key diagnostic tool due to its safety, cost-effectiveness, and real-time capability, making it indispensable across a wide range of medical applications. By incorporating advanced deep learning (DL) and artificial intelligence (AI) methodologies, this thesis research aims to enhance the diagnostic accuracy and efficiency of existing US imaging analyses focusing on discriminative DL models. It also seeks to unlock new applications particularly considering generative DL models, thereby broadening the spectrum of ultrasound's utility in medical diagnostics.

The development of robust DL models and data processing pipelines that address the inherent limitations of current techniques for B-mode ultrasound image analysis is a central focus. Three distinct clinical contexts that each present their specific challenges are confronted and novel solutions are proposed. In particular, B-mode ultrasound imaging and DL applications are affected by issues such as operator bias, limited data availability, and the generalizability and explainability of AI solutions to multi-organ segmentation and classification tasks.

The first challenge of operator bias is tackled through the development of a novel method for training DL models for the clinical task of measuring Common Carotid Intima-Media Thickness (cIMT). Measuring cIMT with manual annotations results in suboptimal risk stratificatification using this biomarker and training DL models directly on manual annotations transfer the operator bias to the models predictions. The proposed method improves the segmentation accuracy of the DL model, using a hybrid of manual and computerized annotations to guide network's training. Moreover, a custom, task specific modification to Dice loss is proposed. These methods improve the segmentation performance of the DL model when tested in an external multi-center dataset achieving better correlation than manual operator when compared with the average of multiple readers.

The second issue of limited data availability is confronted through the development of a method for the automatic measurement of the Optic Nerve diameter (OND) and Optic Nerve sheath diameter (ONSD). To overcome the challenges brought by having a dataset with low numerosity and high variability, an innovative ensemble on UNet models was developed, validating the increase in segmentation performance with ONS and ONSD clinical endpoints.

The challenge of generalizability and explainability of DL solutions to multi-organ segmentation and classification tasks is addressed through the proposal of an end-to-end pipeline for the classification, segmentation, and characterization of muscle US images. This chapter expands current literature enabling the automatic classification and segmentation of up to 16 muscle groups while previous works typically focus only on a few muscle groups. A first of its kind radiomics-based approach using texture analysis of the segmented muscle and a machine learning model was developed to reproduce the Heckmatt score, a visual assessment scale of muscle condition. Explainability of the AI method was incorporated evaluating feature importance. The developed pipeline is hence able to provide an explainable and reproducible way for muscle analysis that improves manual scoring by quantitatively explaining a categorical grade.

Challenges encountered in earlier discussed applications, such as the variability in manual tracings of the Intima Media Complex segmentation, low signal-to-noise ratio in Optic Nerve Diameter (OND) measurements, and the complexity of segmenting multi-muscular B-mode images, underscore the limitations of discriminative models. These challenges arise from factors like underrepresentation in datasets, imprecise annotations, and suboptimal loss function selections, which collectively diminish model precision.

In order to take a step forward compared to discriminative models and address some of the limitations of these DL models in image analysis, generative models are explored. Generative models offer transformative solutions to these issues by enhancing image quality through denoising, standardization, and augmentation. Moreover, generative models provide the ability to create

realistic medical images that can be used to validate discriminative models and develop simulator training tools for physicians. The final focus of this thesis is the development and proposal of a modular pipeline for medical image synthesis. This methodology proposes breaking down the image generation process into specific sub-tasks, enhancing both model training efficiency and control. Application of this pipeline to B-mode thyroid ultrasound imaging and digital pathology is proposed. The developed framework achieves realistic image generation correctly representing both healthy and pathological structures, and can be leveraged for decision support, as well as for training physicians and refining deep learning models.

In conclusion, this thesis captures the evolution of deep learning applications in medical B-mode ultrasound imaging, highlighting novel methodologies and the creation of a modular pipeline for image synthesis, aiming to improve diagnostics, physician training, and to extend the application of AI in enhancing patient care and medical education.

Table of Contents

SUMMARY **4**

1. INTRODUCTION **10**

1.1. MOTIVATION 10
1.2. STRUCTURE OF THE THESIS 11

2. BACKGROUND **12**

2.1. ULTRASOUND IMAGING 12
2.1.1. AI applications in B-mode Ultrasound *15*
2.1.2. Challenges for AI in Ultrasound *16*
2.2. DISCRIMINATIVE MODELS 16
2.3. GENERATIVE MODELS 20

3. INTIMA-MEDIA THICKNESS ESTIMATION AND BIOMARKER ANALYSIS **24**

3.1. INTRODUCTION 24
3.2. DATASETS AND CLINICAL STUDIES DESCRIPTION 26
3.2.1. CUBSclinical *26*
3.2.2. CUBStechnical *28*
3.2.3. Multicenter *29*
3.3. IMT AS A BIOMARKER – CLINICAL STUDY 30
3.4. HYBRID GROUND TRUTH 32
3.5. COMMON DICE LOSS 37
3.6. DISCUSSION 39

4. OPTIC NERVE CHARACTERIZATION IN TRANSORBITAL ULTRASOUND IMAGES ON A MULTICENTER DATA SET 41

4.1. INTRODUCTION 41
4.2. DATABASE DESCRIPTION 42
4.3. MANUAL AND AUTOMATED OND AND ONSD MEASUREMENT 43
4.4. MODEL TRAINING AND INFERENCE 46
4.5. MODEL EVALUATION 47
4.6. RESULTS 48
4.7. DISCUSSION 52

5. DEEP LEARNING AND RADIOMICS FOR QUANTITATIVE MUSCLE ULTRASOUND ANALYSIS **54**

5.1. INTRODUCTION 54
5.2. STATE OF THE ART – AUTOMATIC SEGMENTATION 55
5.3. DATASET DESCRIPTION 56
5.4. SEGMENTATION MODEL DESCRIPTION 58
5.5. PIPELINE DEVELOPMENT AND EVALUATION 59
5.6. RESULTS 62
5.7. DISCUSSION 67

6. A MODULAR APPROACH FOR MEDICAL IMAGE SIMULATION UNCOUPLING SEMANTIC AND STYLISTIC INFORMATION IN A LIMITED DATA ENVIRONMENT **69**

6.1. INTRODUCTION 69
6.2. RELATED WORK 71
6.3. THYROID ULTRASOUND 72
6.4. DATASET DESCRIPTION 72
6.5. MODEL DEVELOPMENT 73
6.6. RESULTS 76

6.7. GENERATIVE MODULAR PIPELINE FOR DIGITAL PATHOLOGY 80
6.7.1. Clinical application 80
6.7.2. Dataset description and preprocessing 80
6.7.3. Methods 83
6.7.4. Results 83
6.8. DISCUSSION 86
7. CONCLUSION 87

List of Figures

Figure 1 Example schematics for medical image analysis tasks applying Deep Learning networks 19
Figure 2 Example of GANs architectures. A) Original GAN. B) Pix2Pix GAN. C) CycleGAN. D) SPADE 22
Figure 3 Vascular structure of the carotid arteries. Illustrations of a healthy artery and an artery in the presence of a plaque. 24
Figure 4 Example of images from each center (A-D) in-vivo images. (E) FAMUS simulated images. (A) Munich, Germany; (B) Pisa, Italy; (C) Porto, Portugal; (D) Torino, Italy; and (E) Toronto, Canada. 29
Figure 5 Kaplan–Meier cardiovascular event analysis plots, considering four CIMT groups: <650 mm, between 650 and 750 mm, between 750 and 850 mm and >850 mm. A1, A2 and A3 = manual segmentations of analysts 1, 2 and 3, respectively. A1' = segmentations of analyst 1, traced 1 month after A1; UCYCY = computerized method from University of Cyprus; TUMDE = method from Technische Universität München; CNRIT = method from Consiglio Nazionale delle Ricerche; INESCTECPT = method from INESC Technology and Science; POLITOIT = method from Politecnico di Torino. 31
Figure 6 Illustration of the Hybrid GT construction. The profiles are A1 (red), A1s (blue dashed), TUM (yellow), Hybrid GT (green dotted). In panel A there are the three available profiles (two manual, A1 and A1s, and one computerized, TUM). In panel B, the condition where only one profile is available is shown. In panel C, two profiles are available. In panel D, three profiles are again available. 34
Figure 7 Training (top row) pipeline and inference (bottom row) pipeline. 34
Figure 8 Bland-Altman plots comparing inter-operator variability and UNetHybridGT results on the test set. Manual measurements are referenced as GT1, GT2, GT3. UNetHybridGT is referenced as UNet. RPC: reproducibility coefficient (1.96*standard deviation of the errors). The horizontal lines define the mean error (black solid line), and upper and lower limits (dashed red lines, ± 1.96 SD). In bold the mean differences with a significant statistical difference (p<0.01). 36
Figure 9 Examples of images available in our dataset from different centers and machines. A) Homburg, Germany (MyLab Gold 30). B) Turin, Italy (Esaote MyLab). C) Novara, Italy (Toshiba Medical System Aplio 300). D) GE-Vivid7. 43
Figure 10 In panel A is displayed the image cropping process. Panel B illustrates the usage of the two GUIs to extract masks and diameters. Panel C shows UNet training and inference steps. 44
Figure 11 A) Custom User Interface for diameter measurement. On the left panel, the operator can measure OND and ONSD by tracking two lines. On the right panel, there is a guide that highlights the 3 mm zone. B) ImageLabeler User Interface for ONS segmentation. 45
Figure 12 A) Depiction of the optic nerve and its meningeal sheath. B) OND (blue arrow) and ONSD (red arrow) measured 3mm behind the papilla. C) Measurements of the OND and ONSD in a healthy subject. D) Measurements for a pathological subject. 45
Figure 13 A) Binarized SoftMax outputs from 8 separately trained models. B) Post-processed SoftMax. C) Averaging of the 8 SoftMax results in a probability map. D) Post-processing of the probability map (thresholding at 0.6 and morphological operations). 47
Figure 14 Areas considered during Signal and Noise Power calculation. A) Signal power (green label) is calculated in the ONS area, while Noise power (red label) is in the ON area. B) Signal Power is calculated in the ONS area and Noise power in the retrobulbar fat. 48
Figure 15 A) Boxplots showing the Dice scores for the 8 runs and the ones resulting from the Ensemble of the SoftMax. B) Boxplots of Hausdorff distances in pixels between the ONS contours of the ground truth and the automated masks for the 8 runs and the Ensemble. 50
Figure 16 Dice scores for the subsets when that machine was used as the test set. First column is the union of the four runs. B) Hausdorff distances. Mach 1 (Esaote), Mach 2 (Esaote), Mach 3 (Aplio300 and AplioXG), Mach 4 (Vivid7). 50
Figure 17 A) Bland-Altman and Correlation plots. B) ONSD. The red dashed lines are at 1.96 SD. rho is calculated using Pearson's correlation. A1: manual measurements obtained by the most experienced operator. 51
Figure 18 A) Boxplots visualizing Dice Score and Hausdorff Distance distributions for distinct categories of SNR. B) These Boxplots visualize SNR distribution depending on the correctness of the OND measurement (GT ± 0.2 mm). SNR is calculated first from the ON (left), then from the ON sheaths to their peripherical area(right). C) SNR distribution depending on the correctness of ONSD measurement. 51

Figure 19 Dataset muscle type distribution after selection. Zygomaticus presents a higher count since we grouped Zygomaticus major and Zygomaticus minor images under physician evaluation give their similar appearance and physiology. 58
Figure 20 Segmentation framework used in this work. (A) Features are extracted by using a Swin transformer as backbone. (B) Multiscale features aggregation is performed by the UPerNet (Unified Perceptual Parsing for Scene Understanding). (C) Overall architecture of the KNet. During the inference process, the network generates a softmax output that indicates the probability of each pixel belonging to one of the muscle or background classes. 59
Figure 21 Mean Intersection over Union for different runs on the validation set. 60
Figure 22 Project pipeline. A) Multiple repeated bilateral acquisitions. In the figure are shown examples from Rectus Abdominis, Biceps Brachii, Tibialis Anterior, Gastrocnemius Medialis B) The segmentation network performs muscle classification and segmentation, features are extracted from the CSA and from the area under the muscle using PyRadiomics. C) Feature dimensionality is reduced considering spatial robustness and correlation. Remaining features are averaged if multiple acquisitions are present. D) The ordinal regression model predicts probability of Heckmatt class. 61
Figure 23 Segmentation results for different muscle types for the same subject. 63
Figure 24 Normalized confusion matrix (whole dataset N=24756). 65
Figure 25 Boxplots for IoU by fold. 66
Figure 26 A) Normalized confusion matrix for prediction of Heckmatt grade. B) (top right) ROC One-vs-Rest approach. C) (bottom right) ROC One-vs-One approach. 66
Figure 27 Pipeline for controllable image generation. 70
Figure 28 Process for determining best crop for ultrasound image. 73
Figure 29 DP-GAN Generator. The first pyramid (orange) comprises spatially-adaptive normalization (SPADE) blocks and upsampling, which transforms a downsampled version of the input into an image. Meanwhile, the secondary pyramid (blue) is responsible for refining the conditioning of the SPADE blocks in the orange pyramid. Moreover, the lowest resolution layer output in the yellow pyramid undergoes upsampling and is then integrated (indicated by C) with the features from preceding layers. 74
Figure 30 DP-GAN discriminator. 76
Figure 31 Losses and FID during training. 77
Figure 32 On the left Precision and Recall calculated using the radiomics features at the different timepoints. On the right, the plot of the two principal components of the PCA computed on the radiomics features in the image area at 4 different timepoints. 77
Figure 33 Examples of synthesized thyroid ultrasound images. In the segmentation map: black (external background), gray (internal background), yellow (trachea), cyan (thyroid), magenta (vessels), green (hyper echogenic structures), red (nodule), blue (skin and derma), orange (muscle) 78
Figure 34 On the left Precision and Recall calculated using the radiomics features at the different timepoints for each label. On the right, the plot of the two principal components of the LDA computed on the radiomics features in each label at 4 different timepoints. 79
Figure 35 Intersection over Union of trained network at each checkpoint for each label. 79
Figure 36 Different readers evaluations of the same WSI. 81
Figure 37 Example of real image and the automatically extracted instance map and segmentation map 82
Figure 38 Qualitative result of the simulation of realistic segmentation maps. First row, real images. Second row, generated images. 84
Figure 39 Qualitative results of the histopathological image synthesis using SPADE from segmentation maps. 85

1. Introduction

1.1. Motivation

Medical imaging stands at a key point where technology and healthcare converge, aiming to evolve how we diagnose, treat, and understand human diseases. In the healthcare field, integration of computerized methods in current medical pipelines is not just a technological goal but a necessity due to the need for improved patient outcomes, more efficient healthcare workflows, and the limitations of human assessment. Additionally, there is a pressing need to reduce costs given the scarcity of medical personnel, which not only pressures the healthcare system but also limits access to high-quality, personalized healthcare for a broader population. Addressing this challenge is crucial for advancing healthcare equity and efficiency.

In this thesis, particular attention will be put on ultrasound imaging. Of all the imaging modalities, ultrasound is often one of the first on which the medical personnel rely to assess a patient's condition. Given its lower cost than imaging modalities like CT or MRI and its portability, it can be applied in many clinical scenarios from bedside care to ambulatory analysis, and even in critical care. The acquisition and interpretation of ultrasound images is also one of the hardest among medical images and requires trained operators. This difficulty in the interpretation of ultrasound images is also linked to a high variability in human assessment especially when performing anatomical measurements. Therefore, it’s evident how reliable and reproducible automated systems can help in the standardization of ultrasound image analysis and could improve large quantitative studies.

The path to integration of computerized methods in medical imaging systems and medical image evaluation pipelines is not without challenges. These include the rights to data privacy, the need for algorithmic transparency, and integration into existing healthcare workflows. Deep learning (DL) and more broadly artificial intelligence (AI) hold the potential to take on these challenges and more. They might also help to reveal patterns in medical data that are invisible to the human eye, offering clues for early disease detection, treatment outcome predictions, and new insights into pathology that could transform patient care. With the constant increase in data availability from new imaging techniques, volumetric analyses, improvements in image resolutions, and the pervasiveness of wearable sensors the dimensionality and complexity of data are only going up. To manage this level of complexity, computerized medical image analysis methods must be focused on helping physicians navigate the vast amount of data available. The synergy of computerized methods and human evaluation could therefore mean reducing repetitive and error-prone workload to physicians while granting patients more precise, reliable, and customized cures.

In its computer vision applications, deep learning involves training neural networks on large datasets to recognize patterns and obtain data-driven results. The effectiveness of DL models is heavily reliant on the availability of large, annotated datasets, which serve as the foundation for optimizing a cost function that guides the training of the network. These datasets enable the models to learn from a vast array of examples, improving their ability to make accurate predictions in medical images. Even when in the presence of large datasets another challenge lies in the variability of human interpretation of medical images, where one expert's observation might be another's oversight, leading to inconsistent assessment. This variability in turn reflects in a lowered accuracy of automated systems. Since DL systems learn to mimic human decisions and human behavior, a ground truth derived from unreliable human annotations will result in an augmented uncertainty of the system outputs, especially when prompted with borderline cases. This highlights the need for open and reproducible systems and open datasets that can be validated by multiple research groups. Moreover,

standardized, reliable benchmarks and ground truths are essential to guide automated medical systems toward more reliable results.

With this background, this thesis has the aim of advancing the field of DL applied to medical images by improving existing tools and pipelines for the analysis of medical images. Some biomarkers like common Intima-Media Thickness (cIMT) or Optic Nerve Diameter (OND) are limited in their ability to stratify the risk of the population by the uncertainty in the human evaluation or limitations in image quality. In this research, new methods for the segmentation of the Intima-Media complex and the optic nerve will be investigated to eventually improve cIMT and OND measurement. Moreover, some clinical pipelines use human scores to evaluate disease progression like the Heckmatt score for muscle condition assessment. This leads to qualitative scores that have limitations over quantitative evaluation. In this thesis, this limitation will be approached by developing an automatic scoring pipeline for muscles in B-mode images that could assist the physician in the evaluation of the muscle. While advancing DL-based analysis tools can improve the applicability of such models, some limitations like domain shift, lack of standardization in imaging pipelines, and uncertainty in human assessment still represent a burden on these models, thus requiring the development of new solutions.

In recent years, deep learning models that are not limited to providing a discriminatory output on medical data but can also generate synthetic data have strongly impacted the scientific community. These generative models showed great potential to produce synthetic images, even in medical settings where data is scarce or where annotations have been difficult to obtain. By synthesizing high-quality, realistic medical images, generative models provide an additional resource for training discriminative deep learning algorithms. This approach also facilitates the testing and refinement of these models under controlled conditions, ensuring that they perform well across a range of scenarios in actual clinical practice. Eventually, this thesis proposes a modular approach for image synthesis that untangles the generation of the semantic content of the image and its stylistic components. This method enables the generation of realistic medical images that can be used in turn for the training od physicians or DL models.

1.2. Structure of the thesis

Following this introduction, Chapter 2 “Background” is dedicated to introducing the models and imaging modality that will be explored during the manuscript.

The rest of the thesis is divided into chapters that will present three different medical image analysis applications where the optimization of the DL system has been approached from three different perspectives. Chapter 3 is about vascular ultrasound imaging and the estimation of Intima Media Thickness through the segmentation of the Intima Media Complex. This chapter will focus on the training strategies of Deep Learning segmentation models and on finding the best loss function to model the specific segmentation task. Chapter 4 describes a segmentation method for the Optic Nerve diameter measurement in transorbital B-mode ultrasound images. This chapter describes the challenges of having a multicenter dataset with low numerosity. The proposed solution is a method that merges rule-based and deep learning methods with an ensemble of different networks. The third image analysis application is described in Chapter 5 where, exploiting a large dataset, a pipeline that includes classification, segmentation, and tissue characterization for musculoskeletal ultrasound imaging is proposed. Finally, Chapter 6 stems from the limitations of DL based system for image analysis that will be highlighted in previous chapters and describes a modular pipeline for the controllable synthesis of medical images that could potentially be applied for several downstream tasks. Chapter 7 concludes the thesis with general remarks and consideration.

2. Background

2.1. Ultrasound imaging

Ultrasound imaging offers a non-invasive, real-time view into the human body. Its utility spans across various clinical fields, from cardiology to obstetrics, to musculoskeletal and vascular analysis, providing essential data for the diagnosis and monitoring of numerous conditions. Unlike other imaging modalities such as computed tomography (CT) ultrasound does not employ ionizing radiation, making it a safer alternative, especially for vulnerable population. Its portability and cost-effectiveness further amplify its utility, particularly in low-resource settings or bedside diagnostics. Ultrasound enables the visualization of soft tissue structures, fluid movement, and organ function with remarkable detail. However, its efficacy is not without limitations. Operator dependency, limited penetration in obese patients, and the inherent challenge of interpreting echogenic patterns pose significant challenges. Additionally, the quality of ultrasound imaging is highly contingent on the skill of the operator, making standardized training crucial[1].

Ultrasound image formation starts with the generation of sound waves by a transducer, which converts electrical energy into mechanical energy through the piezoelectric effect. These sound waves, characterized by their frequency (f) and wavelength (λ), propagate through the body with a speed (c) determined by the medium's properties[2]. The relationship between speed, frequency, and wavelength is given by the equation:

$$c = \lambda f \tag{1}$$

where the speed of sound in soft tissues averages 1540 m/s and depends on the density (ρ) and stiffness (K) of the tissue as in $c = \sqrt{\frac{K}{\rho}}$. This value is the average of measurements obtained from healthy soft tissue. While this value is representative of most tissues, aerated lung, bone and fat have propagation velocities that differ significantly from 1540 m/s. This approximation can thus bring errors in measurements or be the cause of artifacts in the image. As sound waves travel through biological tissues, their behavior is influenced by variations in acoustic impedance (Z), a property that quantifies the resistance encountered by the wave[2]. Acoustic impedance is defined as:

$$Z = \rho c \tag{2}$$

When a wave encounters a boundary between two media with different impedances (Z_1 and Z_2), part of the wave is reflected back to the transducer, while the remainder is transmitted or refracted into the second medium. The intensity of the reflected wave is governed by the reflection coefficient (R):

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right) \tag{3}$$

Refraction, or the bending of the wave path, occurs according to Snell's law, altering the direction of wave propagation based on the sound speed differential between tissues:

$$\frac{\sin\theta_i}{\sin\theta_r} = \frac{c_2}{c_1} \tag{4}$$

As ultrasound waves penetrate deeper, they undergo attenuation, a reduction in intensity and amplitude due to absorption, scattering, and reflection. The attenuation coefficient (α), which increases with frequency, affects the depth of penetration and image quality. The cumulative attenuation over a distance (d) can be expressed as:

$$A = \alpha f d \tag{5}$$

The transducer receives the reflected waves, converting them back into electrical signals. These signals, representing varying levels of echo return from different tissues, are processed to create an image. The distance to reflecting structures is determined by measuring the time delay (t) of the echoes, with the distance (d) calculated as:

$$d = \frac{1}{2} ct \tag{6}$$

This calculation assumes the sound wave's round trip from the transducer to the tissue interface and back[2].

As both the originator of sound waves and the detector of their echoes, the probe plays a triple role as a transmitter, transducer, and receiver. Ultrasound imaging employs a range of probes, each optimized for specific clinical applications. Linear probes have a straight array of piezoelectric elements producing high-resolution images of superficial structures. They operate at higher frequencies and are indispensable in musculoskeletal, small parts, and vascular imaging due to their capacity for detailed visualization. In contrast, convex probes feature a curved array that expands the field of view, making them particularly useful for abdominal, obstetric, and gynecological imaging. Their lower frequency range allows for deeper tissue penetration, capturing broader anatomical areas. Phased array probes have a small footprint and can steer the ultrasound beam through precise timing variations, are essential in cardiology. Their design facilitates imaging between ribs and other tight spaces, providing a deep field of view from a small acoustic window. Endocavitary probes are designed for insertion into body cavities, offering high-resolution images of internal structures near the cavity walls. They find their application in transrectal and transvaginal imaging, allowing close-up views with higher frequency emissions. Further specialized are intravascular ultrasound (IVUS) probes, miniature devices mounted on catheters for detailed imaging of blood vessels from within, aiding in the assessment of vascular diseases and interventions. Lastly, the advancement in probe technology has introduced 3D/4D probes capable of acquiring volumetric data[1]. These are particularly transformative in obstetrics for three-dimensional fetal imaging and in cardiology for comprehensive heart structure analyses. In its transmitting phase, the probe emits short bursts of ultrasound waves into the body. Upon voltage application, the piezoelectric crystals undergo deformation, emitting ultrasound pulses into the body. These pulses are defined by their initial amplitude, frequency, and duration. The **pulse repetition frequency (PRF),** which measures the number of pulses emitted per second, is a crucial parameter in ultrasound imaging. PRF influences both the depth of penetration and the frame rate of the imaging process. A higher PRF allows for faster image acquisition but limits the maximum imaging depth due to the need for pulses to return before the next emission. Conversely, a lower PRF enables deeper penetration at the cost of image acquisition speed. The probe then shifts to its receiving phase, capturing the reflected sound waves. These returning echoes are converted back into electrical signals by the piezoelectric effect, forming the basis for the ultrasound image[3].

Image acquisition parameters control the quality of ultrasound images. **Frequency** and **gain** are among the most significant parameters, with direct impact on image resolution and contrast. The **frequency**

of the ultrasound pulse determines the resolution and depth of penetration. Higher frequencies provide better resolution due to their shorter wavelengths, allowing for more detailed visualization of superficial structures. However, they also attenuate more rapidly, limiting their depth of penetration. Lower frequencies, in contrast, penetrate deeper into the body but provide lower resolution. **Gain** setting control the amplification of the returning echoes to enhance the visibility of structures. This function is critical given the vast dynamic range of reflected signals, which can vary by a factor of up to $1{:}10^{12}$ in clinical settings, translating to a dynamic range of up to 120 dB. While the amplifiers in ultrasound machines are designed to handle this extensive range of voltages, the limitation arises from the gray-scale displays, which can only showcase a signal intensity range of approximately 35 to 60 dB. To address this discrepancy, the receiver compresses the wide range of amplitudes into a manageable spectrum that can be effectively displayed. This compression is achieved by selectively amplifying weaker signals, ensuring that details from both high and low-amplitude signals are preserved and made visible on the display. The process of compression and remapping is essential for adapting the dynamic range of the backscattered signal intensity to the much narrower dynamic range of the display. **Time Gain Compensation (TGC)** is another critical parameter that adjusts the gain dynamically to compensate for the increased attenuation of sound waves at greater depths. By increasing the gain for echoes returning from deeper within the body, TGC ensures uniform brightness across the image. The depth setting determines the range of depth displayed on the screen, affecting both the field of view and the frame rate. A deeper setting provides a broader view of internal structures but can reduce the frame rate due to the longer time required for sound waves to travel to the target depth and back. The **focus**, adjustable on many ultrasound systems, enhances the resolution at a specific depth, improving the clarity of structures at the targeted focal zone while potentially reducing it outside this area. The lateral resolution is closely related to focus settings. In ultrasound imaging, the main lobe is the central part of the sound beam emitted by the transducer, where the intensity of the sound energy is highest. The width of the main lobe plays a critical role in determining the lateral resolution of the ultrasound image. A narrower main lobe leads to better lateral resolution because it can differentiate between two points that are closer together. Focusing the ultrasound beam, either electronically or mechanically with a lens, effectively narrows the main lobe at the focal zone, enhancing lateral resolution at that specific depth. However, alongside the main lobe, side lobes are generated. These are areas of lower energy that extend laterally from the main beam path. While much weaker than the main lobe, side lobes can still detect and reflect sound waves from tissues or structures that are not directly in the main beam's path. This can lead to artifacts or the misrepresentation of echoes, as signals received from the side lobes may be interpreted as coming from the main beam's focus. Such artifacts can complicate the interpretation of ultrasound images by introducing false or misleading information[2]. Advanced imaging techniques have been developed to mitigate the impact of side lobes and thus improve image quality. Harmonic imaging takes advantage of the fact that tissues generate harmonic signals (multiples of the transmitting frequency) when subjected to ultrasound waves. These harmonic signals can be more focused than the fundamental frequency, reducing the relative intensity of side lobes compared to the main lobe. Additionally, adaptive algorithms can dynamically adjust the beamforming process to suppress side lobes, further enhancing lateral resolution and reducing artifacts. Typical values of lateral resolution vary depending on the frequency of the ultrasound wave and the focusing capabilities of the transducer. In general, lateral resolution can range from less than 1 mm in the near field and focal zone to several millimeters in the far field. For instance, at a depth of 5 to 15 cm, which is common for many diagnostic applications, the lateral resolution may be approximately 1-3 mm. This resolution allows for the detailed visualization of small structures, such as vessels and nerves, but is also dependent on the effective management of side lobes to prevent image degradation[2].

Artifacts in ultrasound imaging, such as shadowing, mirroring, and refraction errors, present significant challenges to both manual and automatic analysis. These artifacts arise from the complex interactions of ultrasound waves with tissues and are influenced by the parameters and limitations already discussed. **Shadowing**, for instance, occurs when a highly attenuating structure blocks the path of the ultrasound waves, leaving an area devoid of echoes behind it. **Mirroring** can create duplicate images of structures due to reflections from strong interfaces. Refraction errors distort the true location of structures because of the bending of sound waves as they pass through tissues with different speeds. **Speckle**, a granular noise appearing because of the interference of sound waves within structures in the human body of similar size compared to the wavelength, complicates image analysis reducing contrast resolution. While often considered detrimental, speckle can also provide information on tissue characteristics, challenging the balance between reducing noise and preserving diagnostic information[1].

These artifacts and limitations in image quality make automatic analysis of ultrasound images more challenging. Algorithms designed to interpret ultrasound images must account for the variability introduced by these factors, necessitating sophisticated image processing techniques to differentiate between true anatomical structures and artifacts. The need for accurate, automated analysis is growing, driven by the expanding role of ultrasound in diagnostics and the increasing volume of imaging studies. However, the inherently operator-dependent nature of ultrasound imaging, combined with the diverse range of artifacts and limitations, underscores the complexity of developing robust automatic analysis tools.

In this thesis the focus will be on the analysis of B-mode ultrasound images for the extraction of information about the morphology and characteristics of the underlying tissues.

2.1.1. AI applications in B-mode Ultrasound

AI applications in US image analysis focus predominantly on three main tasks: detection, classification, and segmentation. By analyzing ultrasound images, AI systems can offer insight that were previously dependent on the subjective interpretations of the operators. While US are a versatile imaging tool and can be applied to investigate different organs and clinical scenarios, the interpretation of US images required expert operators. Moreover, imaging the different parts of the human body is linked with unique challenges and requirements. In thyroid studies, AI enhances the detection and accurate diagnosis of thyroid nodules, significantly aiding in the reliable characterization of thyroid diseases. This technological advancement supports clinicians by offering precise segmentation of nodular boundaries, which is crucial for deciding the most appropriate management strategies [4]. In breast imaging, AI models detect and classify breast masses, differentiating benign from malignant lesions. This capability is critical as it aids in making informed decisions about treatment options, potentially reducing unnecessary interventions [5]. The liver, a vital organ often assessed for fibrosis and malignancies, benefits greatly from AI applications in abdominal ultrasound. AI-driven models assist in assessing liver lesions and accurately predicting fibrosis stages, which are essential for guiding biopsies and assessing chronic liver disease [6]. In obstetrics and gynecology, AI transforms fetal imaging by automating the detection of fetal standard planes thus adding the detection of anomalies and enhancing the accuracy of biometric measurements. These advanced AI models support obstetricians by predicting fetal health conditions, thus facilitating crucial prenatal care decisions [7]. Cardiovascular imaging also sees significant improvements with AI, especially in echocardiography where automated cardiac chambers segmentation and functional quantification provide cardiologists with essential tools for diagnosing and assessing heart conditions [8]. In musculoskeletal imaging, AI aids in diagnosing disorders by providing precise segmentation of muscle, thus enabling faster and reproducible extraction of texture features. This allows for better assessment of conditions such as

inflammatory muscle diseases or neuropathies, enabling targeted treatments and improved monitoring of therapeutic responses [9].

Although the focus of this thesis is on B-mode imaging, it is important to underline that US can also be used for functional and dynamic investigations. Some notable examples that benefit from AI are Doppler imaging for assessing blood flow in vessels like the carotid or elastography that study the mechanical behavior of the insonated tissues. In both cases, AI aids both in the acquisition of the signals to have more reliable clinical endpoints and in the interpretation of the results to help physicians in the overall clinical evaluation[10].

These applications expand the utility of US outside the morphological study of the human anatomy making them an invaluable tool in a large number of clinical settings.

2.1.2. Challenges for AI in Ultrasound

AI applications in US have expanded significantly across numerous clinical scenarios. Despite the large number of examples available in the literature some limitations inherent to US imaging limit the adoption of AI based automatic systems to US. Unlike other imaging modalities that provide quantitative data and have standardized acquisition protocols like CT or MRI, in US the quality of the image and its following interpretation are highly dependent on the operator's ability, demanding extensive and specialized training for operators to achieve consistent and reliable outcomes. This subjectivity can lead to inconsistencies, especially in complex cases like fetal ultrasound imaging or cardiac echocardiography, where interpretative errors can affect diagnostic outcomes. This results in an increased uncertainty of the final outcome that is often the target used to train supervised automated systems. Hence the development of AI systems has to consider this limitation in the training phase and when validating the results or comparing different models on the same task.

To improve on this aspect AI algorithms can assist medical personnel by simplifying the operational aspects of ultrasound imaging. By aiding the operator in certain processes such as probe placing or acquisition parameters optimization, AI helps reduce the time and expertise required to conduct examinations, making ultrasound more accessible and easier to use in diverse clinical settings

Some limitation of AI applications in US imaging are directly linked to the physics of ultrasound. Ultrasound cannot penetrate bone or air-filled cavities, which poses challenges in applications like pulmonary imaging or in cases involving obese patients for whom fatty tissue limits sound wave penetration. Directly applying AI systems for detection and image processing in these settings results in limited performance mostly due to the low Signal to Noise Ratio (SNR) of the images. In response to these challenges, AI technologies are increasingly being integrated into ultrasound imaging processes by enhancing image quality, standardizing interpretations, and reducing operator dependency. AI-driven techniques such as machine learning algorithms applied to ultrasound image formation promise significant improvements. These include less blurry images and even the development of novel ultrasound modalities through advanced beamforming and super-resolution techniques. Although these advancements require changes to hardware, they represent a shift towards more sophisticated imaging solutions capable of transforming physical measurements into clearer, more meaningful images [11].

2.2. Discriminative models

Ultrasound imaging and other imaging modalities as well require accurate and reliable algorithm to be interpreted in an automatic way. These algorithms must be robust over image variations and

artifacts. While rule-based methods can partially address these requirements to cope with the variability of real-world data, deep learning based methods are the optimal path.

Deep learning has revolutionized the field of computer vision, particularly in medical imaging, by offering advanced solutions for analyzing, classifying, and interpreting medical images. The required tasks for medical image interpretation and analysis can be broadly classified into four categories: image classification, object detection, and image segmentation [1]. At its core, deep learning utilizes neural networks with multiple layers, enabling the extraction of multiscale features from raw input data. The feature extraction step is a key part of nearly all downstream tasks, thus research on new architectures focused on efficient extraction of features and their reconstruction to achieve the final task. A basic CNN consists of an input layer, multiple hidden layers, and an output layer. The hidden layers include convolutional, pooling, fully connected, activation, and normalization layers. Early CNNs like AlexNet [15] encountered specific problems that impacted their effectiveness and efficiency. One such issue was the problem of vanishing gradients, a phenomenon where the gradients of the loss function decrease exponentially with each additional layer in deep neural networks, making it difficult to update the weights of earlier layers through backpropagation. This issue was particularly pronounced in early CNN architectures, which limited their depth and, consequently, their ability to learn complex features from medical images. The vanishing gradient problem made it challenging to train deeper networks that could capture the hierarchical structure of medical image data, thus hindering the performance of CNNs on more complex image analysis tasks. In response to this problem Residual Networks (ResNets) [16], introduced shortcuts or "skip connections" between layers. These connections allow gradients to flow through the network more effectively during training, enabling the development of much deeper networks without the vanishing gradient problem. ResNets have shown remarkable success in various computer vision tasks by allowing models to learn more complex patterns without a significant increase in computational complexity. DenseNets [17] improved upon ResNets by connecting each layer to every other layer in a feed-forward fashion. This design ensures maximum information flow between layers in the network, significantly reducing the vanishing gradient problem and improving feature reuse, thereby making the networks more parameter-efficient. Following DenseNets, the introduction of Squeeze-and-Excitation Networks (SENet) [18] brought a novel approach to feature recalibration within neural networks. SENets incorporate mechanisms that explicitly model the interdependencies between the channels of convolutional features. By reweighting the features obtained from previous layers, SENets enhance the representational power of the network, allowing for more discriminative feature learning. Attention mechanism brought a lot of innovation in feature extraction techniques, most notably through the development of Transformers [19]. Originally designed for natural language processing tasks, Transformers have been successfully adapted for computer vision, leading to the creation of Vision Transformers (ViTs) [20]. ViTs utilize self-attention mechanisms to weigh the importance of different parts of the input image. This method allows for dynamic feature extraction that can adapt to the specific content of the image, offering a more flexible and powerful approach to understanding complex image structures.

Another critical challenge was the need for regularization to prevent overfitting. Overfitting occurs when a model learns the training data to the extent that it performs poorly on new, unseen data. Given the high dimensionality of image data CNNs were particularly susceptible to overfitting. Techniques such as dropout, normalization, and data augmentation were introduced as regularization methods to mitigate this issue. Dropout [21], for example, involves randomly switching off neuron outputs during the training process, forcing the network to learn more robust features that are not dependent on a small set of neurons. Data augmentation techniques, such as rotating, scaling, or flipping images, were also crucial in increasing the effective size of training datasets, further helping

to combat overfitting. Batch normalization [22], normalizes the inputs to each layer within a network such that it reduces the internal covariate shift, which in turn stabilizes the learning process and allows for higher learning rates. This acceleration of training not only improves efficiency but also contributes to the network's generalization capabilities, further mitigating the overfitting problem. Moreover, the advent of novel regularization strategies, such as label smoothing [23] and mixup [24], has provided new avenues for enhancing model performance. Label smoothing regularizes the model by preventing it from becoming too confident about its predictions, which is achieved by softening the target distributions. On the other hand, mixup trains the model on convex combinations of pairs of examples and their labels, promoting the model's generalization by encouraging it to behave linearly in-between training examples. In parallel with these advancements, the exploration of regularization in the context of unsupervised and semi-supervised learning has opened new frontiers. Techniques like contrastive learning [25] and self-supervised learning [26] leverage unlabeled data to learn powerful representations, reducing reliance on large labeled datasets and thereby addressing overfitting through an expanded and more varied data understanding.

The high dimensionality and non-convexity of the loss functions associated with these networks often led to training difficulties, such as getting trapped in local minima or experiencing slow convergence. Addressing these optimization challenges required the development and adoption of more sophisticated optimization algorithms beyond stochastic gradient descent (SGD), such as Adam [27] and RMSprop, which introduced adaptive learning rates for different parameters, helping to accelerate convergence and improve the training stability of CNNs.

These components are common for all the tasks in medical image analysis. The classification process typically makes use of fully connected layers following the feature extraction layers. These layers, also known as dense layers, are where the actual classification is performed. The network uses the features extracted to make predictions about the class of the input image employing a softmax activation function, which outputs the probabilities of the input image belonging to each of the trained classes. The class with the highest probability is then chosen as the prediction. Clinical application of image classification span across a broad spectrum. For example, deep learning has been applied to the automatic classification of X-ray images [28], distinguishing between normal scans, those showing signs of pneumonia, and COVID-19 infections, showcasing the potential of these technologies to support pandemic response efforts. Similarly, CNNs have been leveraged for the detection of lung nodules in CT scans [29], demonstrating improved diagnostic accuracy and the potential for early detection of lung cancer. Other applications involve skin lesion classification [30], [31], [32], retinopathy detection in fundus images [33], [34], early detection of Alzheimer’s disease [35], optical imaging as in classification of blood cells [36] and digital pathology [37]. DL-based image classification has also applied in ultrasound imaging [38]. Notable applications are thyroid nodule classification [4], [39], breast lesion classification [40], liver cirrhosis [41], and Fetal Facial Standard Plane Recognition [42].

Object detection diverges from classification by not only recognizing what objects are present in an image but also pinpointing their location. This is achieved through the implementation of region proposal algorithms that identify parts of an image where objects are likely to be found, followed by classification and bounding box regression to specify object categories and their precise coordinates, respectively. Two prominent architectures that have been widely applied for object detection are Region-Based Convolutional Neural Networks (R-CNNs) [43] and You Only Look Once (YOLO) [44]. These two approaches, and the methods that improved upon them, are fundamentally different. R-CNNs generate region proposals and then run a classifier on these regions, while YOLO performs object detection in a single regression analysis, directly predicting classes and locations for objects in one go,

significantly speeding up the process. Clinical applications of object detection in medical imaging are broad and impactful. They span across various imaging modalities including X-rays, CT scans, MRI, and ultrasound imaging. For instance, a modified version of YOLO algorithm have been useful in identifying in real-time lung nodules in CT scans [45]. In MRI scans, these algorithms assist in detecting brain tumors [46], stroke lesions [47], and spinal tumors [48] providing critical information for surgical planning and treatment strategies. Similarly, in ultrasound imaging, object detection plays a role in detecting anatomical structures and diagnosing conditions such as fetal abnormalities during pregnancy [49], detecting breast cancer lesions [50], and thyroid nodules [51]. Given the temporal resolution of ultrasound, object detection algorithms have been in ultrasound videos to track anatomical structures like thyroid nodules and surrounding structures [52] and vagus nerve [53].

Image segmentation involves partitioning a digital image into multiple segments or group pixels with a similar semantic content usually defined by the underlying structures in the image. This process is fundamental for clinical applications that require precise organ delineation, tumor segmentation, and any other applications that required the precise study of texture or shape of a particular object. Fully Convolutional Networks (FCN) [54] were among the first to adapt CNN architectures for pixel-wise segmentation, eliminating fully connected layers, used for classification purposes, in favor of convolutional layers to output spatial maps instead of classification scores. In 2015 the development of Unet [55] scored a breakthrough for medical image segmentation and its encoder-decoder structure is still at the base of most of the newer models. Improvements were subsequently made with architectures like UNet++, which redesigned skip connections to exploit multiscale features more efficiently, leading to enhanced segmentation performance across various medical imaging tasks [56].

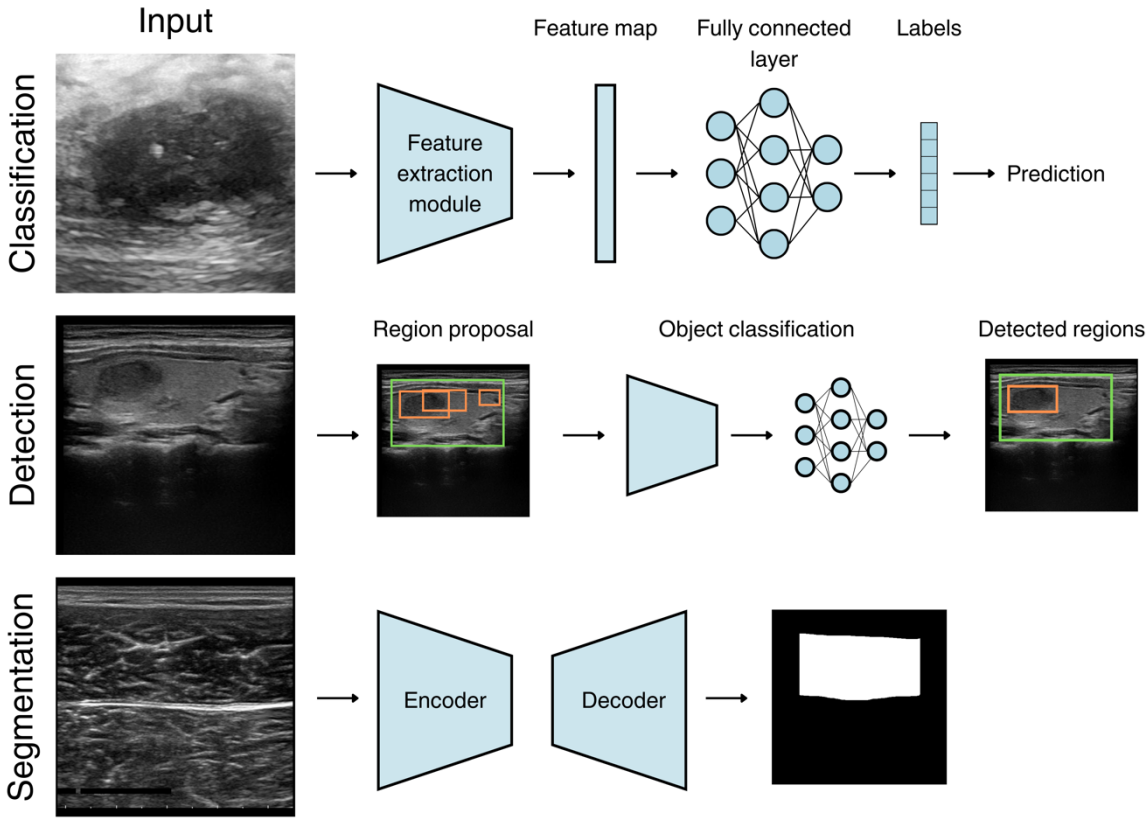


Figure 1 Example schematics for medical image analysis tasks applying Deep Learning networks.

The DeepLab model [57] introduced significant advancements in semantic image segmentation, employing atrous convolution and atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scale improve the localization of object boundaries by combining methods from DCNNs and probabilistic graphical models. More recently, UNEt TRansformers (UNETR), leveraged a transformer as the encoder to capture global multi-scale information effectively. This approach combines the "U-shaped" design's encoder and decoder via skip connections at different resolutions, setting new benchmarks in volumetric (3D) medical image segmentation and achieving state-of-the-art performance on several datasets [58]. Clinical applications of these methods cover all the imaging modalities and several reviews have explored this topic in different ways [59], [60], [61]. Similarly to other task, image segmentation is widely applied to US imaging to delineate structures of interest and performing texture and shape analysis, these applications have been described in [62], [63].

Discriminative models have made a significant impact in the field of medical imaging, demonstrating remarkable results. However, the adoption of DL methods in clinical practice is not widespread. This hesitance is largely due to the black-box nature of neural networks and the challenge posed by real-world data that often differs substantially from the data used during training. To address the black-box issue, the concept of explainable AI (X-AI) [64] was introduced, aiming to shed light on how these models arrive at their conclusions. Techniques like Grad-CAM [65] have become popular, though their effectiveness in explaining model's remains under debate given its susceptibility to adversarial manipulation [66]. Recognizing the limitations of X-AI efforts, recent research has shifted towards quantifying the uncertainty of predictions made by DL models [67], [68]. The idea is that if we can measure how confident a model is in its prediction, it could make these tools more reliable in a clinical setting. In addition to improving interpretability, there has been an effort to enhance the performance of DL models on external datasets. This includes developing new architectures tailored to specific medical imaging tasks, as seen in the previous part, refining loss functions to reflect the complexity of these tasks and incorporating physical principles to guide model predictions more accurately. Efforts to standardize data from multiple centers and apply advanced preprocessing techniques are also crucial for improving model robustness and reducing the impact of out-of-distribution (OOD) data. Despite these advances, the issue of OOD data remains a significant barrier to the routine use of DL in clinical settings, highlighting the need for ongoing innovation in model development and training methodologies.

2.3. Generative models

Generative models extend beyond the capabilities of discriminative models to not just classify or segment existing images but to create new ones. These models enable the synthesis of realistic medical images, data augmentation, and the creation of diverse datasets for training deep learning models. Generative models, particularly through their ability to synthesize new data points, offer a novel approach to addressing the Out-of-Distribution (OOD) problem. One of the most accepted theories about the behavior of generative models is that they learn a mapping from a low dimensional latent space to a high-dimensional data space [69]. By doing so these models can interpolate and extrapolate the underlying data manifold. This capability allows them to interpolate across the data manifold filling the gaps in the training data distribution, potentially improving model robustness against OOD samples by enriching the dataset with diverse examples that were previously underrepresented or missing [70]. In this thesis, one family of generative models will be applied: Generative Adversarial Networks (GANs) [71].

A GAN consists of two neural networks: a generator (G) and a discriminator (D), trained simultaneously through adversarial processes. The generator network aims to synthesize data indistinguishable from

real data, while the discriminator tries to differentiate between real and generated data. The generation process starts from a latent variable represented by the noise vector $p_z(\boldsymbol{z})$. The training objective can be formulated as a min-max game:

$$\min_G \max_D V(D,G) \;=\; \mathbb{E}_{x\sim p_{data}(\boldsymbol{x})}[\log D(x)] \;+\; \mathbb{E}_{z\sim p_z(\boldsymbol{z})}\left[\log\left(1 \;-\; D\big(G(z)\big)\right)\right] \qquad (1)$$

Where x is the real data, z is the noise input to the generator, and p_{data} and p_z are the data and noise distrbutions, respectively. To improve convergence, during training of GANs, the generator and discriminator are updated in an alternate way. While one is updated the other networks is kept fixed. This allows both networks to move towards an optimal point. From the initial implementation in 2014, GANs have seen significant advancements. Improvement in the generated image quality came from different lines of research: some groups focused on working on the theoretical background improving training techniques [72], others worked on the optimizing the objective function like in Wassertein GAN [73] and applying optimal transport theory [74]. Architectural modifications over the initial GAN were explored imposing constraint on the architectural topology like in DCGAN [75]. One of the most important architectural advancement was the introduction of progressive growing of both generator and discriminator during GAN training in ProGAN [76] that allowed generation of fine details at higher resolutions compared to traditional GAN. StyleGAN [77] added an additional mapping network that transforms the input latent code to a traversable mapping to achieve a less entangled latent representation hence getting more control over image features.

The output of a GAN can be controlled by a conditioning variable y that provides more information to the network to generate an output closer to a specific distribution. This condition reflects in the training objective in the following way:

$$\min_G \max_D V(D,G) \;=\; \mathbb{E}_{x\sim p_{data(x)}}[\log D(x|y)] \;+\; \mathbb{E}_{z\sim p_{z(z)}}\left[\log\left(1 \;-\; D\big(G(z|y)\big)\right)\right] \qquad (2)$$

Conditioning in GANs serves to direct the generation process, making it possible to synthesize images with specific attributes or in accordance with certain constraints. This is particularly useful in medical imaging, where generating images that adhere to certain parameters or features can significantly enhance the utility of synthetic data for training and analysis. Conditioning can be achieved by very different inputs like text [78], categorical labels [69], images[79], and feature vectors[80]. One of the most important tasks in conditional generation for medical imaging is the image-to-image generation as in the Pix2Pix architecture[79]. Pix2Pix uses a conditional generative adversarial network to learn a mapping from an input image to an output image in a paired manner, employing a U-Net-based architecture for the generator and a PatchGAN classifier for the discriminator. The L1 loss is incorporated alongside the adversarial loss to ensure the generated images are not only realistic but also closely aligned with the target image on a pixel-by-pixel basis. This characteristic makes Pix2Pix particularly effective for tasks in medical imaging where the precise structure and details of the images are critical. CycleGAN [81] extends image-to-image translation by enabling unpaired training. This model eliminates the need for paired training data, thus broadening the applicability of GANs in medical imaging where paired datasets are scarce or not feasible. CycleGAN utilizes a cycle-consistency loss to ensure that the original image can be recovered from the generated image, thereby preserving key image attributes. Building upon the foundations laid by Pix2Pix, SPADE (Spatially-Adaptive (De)normalization) further advances the concept of conditional generation by introducing a novel normalization technique that makes possible conditioning the output of the GAN on a segmentation map. SPADE blocks modify the traditional batch normalization process by incorporating the segmentation map as an additional input and normalizing the activations in a layer by first applying

a standard normalization process (e.g., batch normalization) and then modulating the normalized output with scaling and shifting parameters derived from the segmentation map. This segmentation map serves as a detailed guide for the generation process, allowing SPADE to produce images that are not only high in quality but also accurately reflect the structure and semantics detailed by the segmentation map. The SPADE normalization adjusts feature maps in a spatially adaptive way, meaning it can modify the style of each area of the image based on the characteristics defined in the segmentation map. This approach is particularly beneficial in medical imaging for generating detailed and realistic images conditioned on precise anatomical and pathological information.

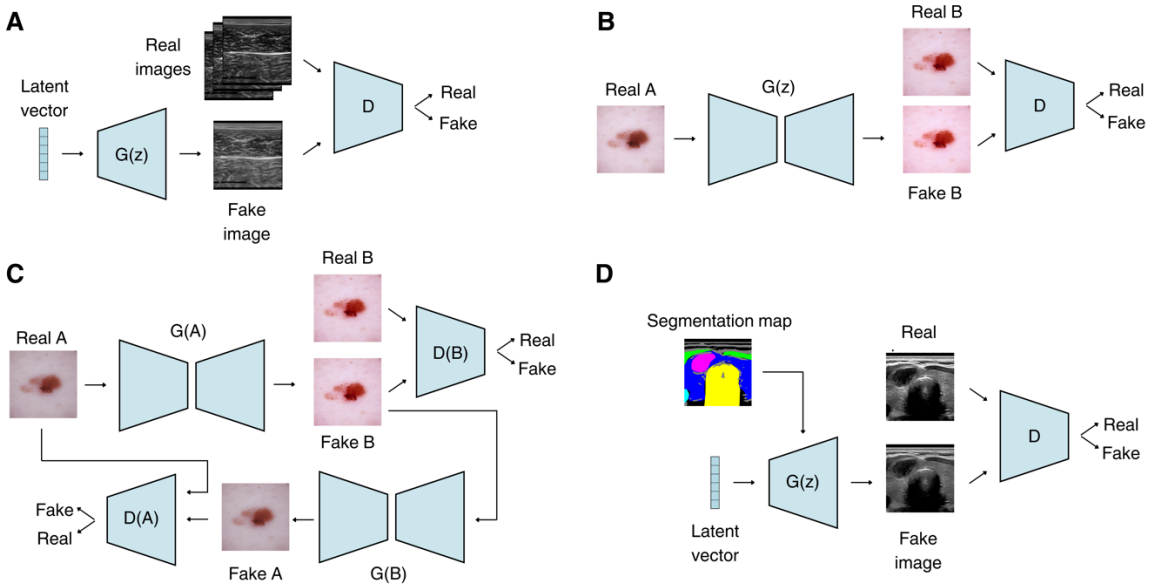


Figure 2 Example of GANs architectures. A) Original GAN. B) Pix2Pix GAN. C) CycleGAN. D) SPADE

GANs can generate high-quality, realistic images, useful for data augmentation and training deep learning models and they have been successfully applied to tasks like image synthesis, super-resolution, and domain adaptation in medical imaging. The clinical applications of SPADE and similar models include generating synthetic yet realistic medical images for training purposes, such as detailed organ segmentation maps being used to generate corresponding anatomical images. This can significantly aid in training medical personnel and developing diagnostic models by providing a rich dataset of anatomically accurate images.

The clinical applications of GANs in medical imaging span across various domains, demonstrating their versatility and significant potential for improving medical diagnostics, training, and research. GANS have been applied for data augmentation in multiple imaging modalities as an intermediate step before segmentation or classification [82], [83], [84]. Ma et al. imposed structure and illumination constraint to the CycleGAN framework to enhance images from different modalities [85]. Image standardization is another field where GAN can provide robust results. Given its high variability in illumination and acquisition condition dermatological images can benefit from GAN standardization as demonstrated in [86]. Digital pathology faces similar challenges hence GAN have their application also in this field [87]. Image to image is the most represented task especially for cross domain applications like multi-contrast MRI [88] or MR to CT translation [89].

While being applied for several tasks, GANs training can be challenging due to issues like mode collapse, where the generator produces limited varieties of samples, or lack of convergence when the min-max game cannot reach an equilibrium point. Therefore, the adversarial training requires careful

selection of hyperparameters. In response to these limitations, the field has seen the development of newer generative models that offer improvements in stability, sample diversity, and ease of training. Variational Autoencoders (VAEs) [90], [91], Normalizing Flows [92], and Diffusion Models [93] are alternative approaches to data generation. VAEs approach the generation task by learning a latent representation of the input data, which is then used to reconstruct the data. This method helps mitigate mode collapse by encouraging exploration of the entire data distribution and provides a more stable training process compared to the adversarial approach of GANs. Despite their notable contributions, VAEs encounter several challenges that limit their efficacy in this field. One of the primary issues is their difficulty in capturing the high-level structures and the broad variability inherent in pathological features within medical images. This limitation significantly affects their utility in accurately detecting anomalies and recognizing pathology from medical datasets.

Normalizing Flows, model the data distribution through a series of invertible transformations, allowing for the exact computation of data likelihoods. This capability enables the generation of high-quality samples and offers a transparent mechanism for model evaluation and interpretation, addressing some of the evaluation challenges faced by GANs. Despite their potential, the requirement for complex, invertible functions introduce significant challenges, especially for high-dimensional data typical in medical imaging.

Diffusion models represent another breakthrough in generative modeling, generating data by reversing a process that gradually adds noise to the original data. This approach has shown exceptional performance in generating high-quality images, including in the medical imaging domain. Diffusion models tackle GANs' issues like mode collapse and training instability by providing a more gradual and controllable generation process. While offering improved stability and sample diversity over GANs, diffusion models' computational demands and slower inference times pose challenges for real-time medical imaging applications.

While each of these newer models brings its own set of advantages, they also come with their own challenges. However, their development signifies ongoing efforts to overcome the limitations of GANs, offering promising directions for enhancing the generation of medical images and furthering the capabilities of computational models in healthcare.

3. Intima-Media Thickness estimation and biomarker analysis

3.1. Introduction

Cardiovascular disease (CVD) stands as the leading cause of death and disability globally. In 2019 alone, the total cases of CVD were estimated at an alarming 523 million, highlighting its significant impact on worldwide health [94], [95]. A critical component of CVD is atherosclerosis, a condition where arterial blood flow is affected, severely limiting the supply of blood to essential organs [96]. Particularly noteworthy is the role of atherosclerosis in the carotid arteries, contributing to approximately a quarter of all stroke cases [97]. These statistics underscore the importance of studying carotid arteries, especially the common carotid artery (CCA) that accounts for most of the manifestations of atherosclerosis.

The CCA, divided into the left and right branches, is a key vessel in the cardiovascular system, primarily responsible for delivering blood to the brain [98]. Maintaining the health of the CCA is, therefore, directly linked to the brain's well-being. The normal and abnormal structures of the vascular system, including the presence of plaques, are depicted in Figure 3.

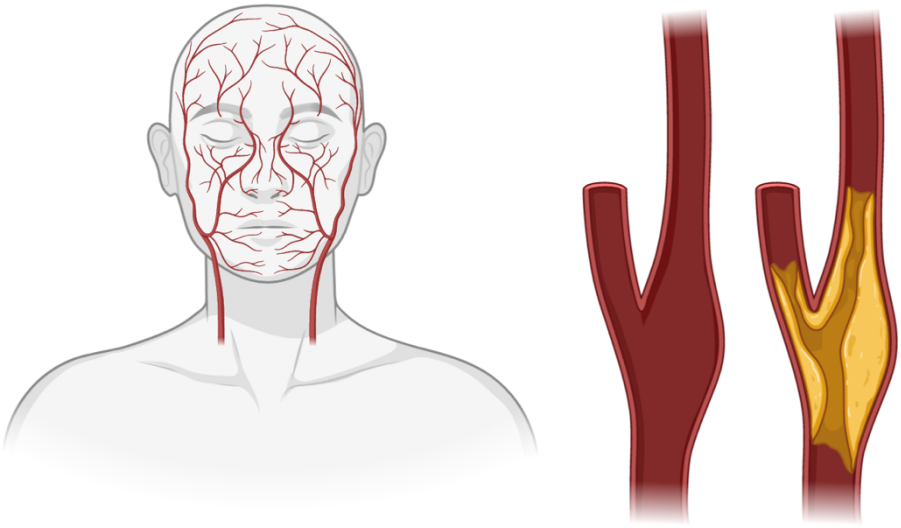


Figure 3 Vascular structure of the carotid arteries. Illustrations of a healthy artery and an artery in the presence of a plaque.

To diagnose conditions like carotid artery stenosis, various imaging techniques are employed, including CT, ultrasound imaging, and MRI. Among these, ultrasound imaging is often preferred for its ability to provide satisfactory resolution while examining the complex structures of the carotid artery [99]. The advancement of medical image analysis, from traditional image processing to modern machine learning and deep learning approaches, has greatly enhanced the automatic segmentation, analysis, and grading of stenosis in carotid artery imaging.

The Common Carotid Intima-Media Thickness (CIMT), measured by delineating the contours of the intima-media complex (IMC), is a well-recognized surrogate marker for atherosclerosis. Utilizing B-mode ultrasound imaging of the CCA, researchers can estimate CIMT. Its quantification is typically performed by segmenting the CCA on the far wall, following the guidelines set out by the Mannheim

consensus [100]. Segmenting the CCA involves determining the position of the lumen-intima (LI) and media-adventitia (MA) anatomical interfaces.

Large epidemiological studies have highlighted the predictive value of CIMT for myocardial infarction or stroke, independent of traditional cardiovascular risk factors [101]. This evidence supports the clinical relevance of CIMT in predicting future CVD events, especially in high-risk individuals. However, the utility of CIMT as a predictive biomarker is not without controversy. Some studies have reported negative results regarding its independent predictive value for cardiovascular events [102]. The additional value of CIMT measurement in risk classification, when combined with existing risk scores, appears to be limited [103]. Furthermore, the association between CIMT progression over time and cardiovascular risk is not always clear, partly due to small changes in vessel walls over time and potential measurement errors [104].

Factors influencing CIMT estimation include the measurement location on the CCA, the ultrasound protocol (single vs. multiple angles acquisition), and equipment setup variables such as transducer central frequency, filters, image gain, depth, and time-gain compensation. These factors necessitate standardization, particularly when using B-mode based systems.

Most large prospective cohort studies on CIMT's predictive value have relied on manual measurements, which can be prone to inter- and intra-analyst variability. To address these limitations, numerous computerized methods have been proposed for extracting the contours of the IMC and measuring CIMT in longitudinal CCA ultrasound images. These methods, which include techniques based on edge operators, active contours (snakes), dynamic programming, adaptive thresholding, fuzzy C-mean clustering, and machine learning approaches, typically offer increased measurement robustness and faster analysis times. Machine learning and deep learning approaches have shown promise in enhancing CIMT measurement accuracy.

Recent studies comparing different computerized CIMT measurement methods on a common large clinical database have shown that computerized segmentation is comparable to skilled analyst's segmentation in terms of correlation with clinical parameters and the ability to predict cardiovascular events [105]. This finding suggests that computerized methods, including advanced ML and DL techniques, can be effectively used interchangeably with manual methods for CIMT quantification and clinical outcome investigation. This advancement holds great promise for standardizing CIMT measurements, thereby enhancing their utility in risk stratification and assessment.

One of the primary challenges in intima-media segmentation arises from the quality of US images. Artifacts in US imaging can obscure the grayscale differences between the LI interface and the lumen, making it difficult to accurately locate the LI and the MA boundaries. Furthermore, the intima-media area often occupies only a small portion of US images, posing a challenge in focusing on this region. Additionally, the IMC may appear intermittently in US images, and different ultrasound equipment brands or parameter settings can result in varying image intensities. These challenges necessitate robust CAD systems that can adapt to these variations.

In response to these challenges, an array of methods has been developed and refined over time. Initial approaches employed traditional techniques like parametric models and snakes [106], [107], which utilized deformable contours for detecting the IMC. Thresholding methods [108] and edge operators [109] were also explored to delineate the IMC boundaries. Dynamic programming [110], which added an augmented dimension for processing 2-D images, marked a significant advancement in this field.

With the advent of machine learning, techniques like Support Vector Machines (SVM) [111], Random Forest (RF) [112], and Fuzzy C-means clustering [113] were introduced. These methods enhanced the segmentation's robustness by updating cluster centroids and membership values based on local pixel

information. The emergence of deep learning brought a paradigm shift in IMC segmentation. UNet and its variants [114], [115], [116] demonstrated exceptional capabilities in handling the segmentation tasks. More advanced models like BSMNet [117] and CSM-Net [118] introduced innovative features like parallel linear structure learning modules and spatial attention mechanisms, significantly enhancing segmentation accuracy. Comprehensive reviews on these advancements [119], [120], [121], [122], [123] have provided detailed insights into the evolution of segmentation techniques, from traditional methods to modern ML and DL approaches. These reviews not only cover the range of applications in cardiovascular imaging but also compare the performance of various methods, thereby highlighting the future directions in this rapidly evolving field.

This chapter shows the research performed on this topic. Firstly, the datasets used to develop the methods and subsequently shared with the scientific community are presented. Then, the clinical study about the clinical significance of IMT as a biomarker is showcased. From the need of better methods for IMT estimation underscored from this section, the following ones will go into detail in the developed methods for IMT estimation. The first research about hybrid ground truth is therefore presented. In this research, to improve the segmentation accuracy of the deep learning model, a synergy of manual and computerized annotations was used to guide the network training. Following this part, the second research about the tuning of Dice Loss for CCA segmentation is presented, demonstrating how focusing the attention of the loss in the correct area of the image benefits the learning of the segmentation network. Finally, a discussion section concludes the chapter.

3.2. Datasets and Clinical studies Description

To develop and validate the approaches described in this chapter three different datasets identified as $CUBS_{clinical}$, $CUBS_{technical}$, and Multicenter were organized and openly shared with the scientific community.

3.2.1. CUBSclinical

The first dataset was collected in the context of CUBS clinical study and published as material of [105] comprising both image data and clinical data for CVD prediction. In this study, 1088 participants were included. It involved acquiring CCA ultrasound images from both sides of the neck, resulting in 2176 images. These images were collected from two distinct groups. The first group consisted of 694 participants from the villages of Pedoulas, Nissou, and Kambos in Cyprus, contributing 1388 images collected between 2003 and 2007. All residents over 40 in these villages were identified through the population list and invited to participate. The second group involved 394 participants from the Hypertension Outpatient Clinic at the University of Pisa, with 788 images acquired between 2011 and 2014. These participants were part of two studies that had received approval from the institutional ethics committee. Unlike the Cyprus study, there was no clinical follow-up planned for the Pisa participants. In both groups, all participants provided written informed consent, which included permission for future use of the images in studies. The image acquisition for all participants adhered to the Mannheim consensus guidelines [100]. Additionally, other clinical parameters for each participant were recorded (as outlined in Table 1).

Parameters	Entire Database	Cyprus	Pisa
N° of participants (images)	1088 (2176)	694 (1388)	394 (788)
Ultrasound instrumentation		Philips (ATL) HDI 5000	MyLab25 (Esaote)
Age, years (mean ± std)	62 ± 11	61 ± 10	64 ± 13
Sex	546 females (50%)	316 females (46%)	226 females (57%)
Smoking [n (packs/years)]	418 (unknown)	260 (36 + 33)	158 (unknown)
Hypertension [n (%)]	582 (53%)	242 (35%)	340 (86%)
Diabetes [n (%)]	167 (15%)	91 (13%)	76 (19%)
BMI (kg/m^2)	28.02 ± 4.48	28.05 ± 4.49	27.97 ± 4.47
Glucose (mg/dL)	104.49 ± 31.31	103.72 ± 27.67	106.06 ± 37.64
Total cholesterol (mg/dL)	218.47 ± 42.16	226.64 ± 42.76	202.61 ± 36.08
HDL (mg/dL)	51.55 ± 14.07	50.35 ± 12.53	53.89 ± 16.44
LDL (mg/dL)	131.48 ± 32.05	136.08 ± 30.64	122.45 ± 32.88
Triglycerides (mg/dL)	146.35 ± 85.93	151.64 ± 90.97	135.99 ± 74.14
Creatinine (mg/dL)	0.93 ± 0.24	0.93 ± 0.25	0.92 ± 0.23
Apolipoprotein A1 (mg/L)	-	1.44 ± 0.24	-
Apolipoprotein B (mg/L)	-	1.20 ± 0.24	-
Follow-up CV events [n (%)]	-	127 (18%)	-

Table 1 Subjects characteristics and clinical parameters in CUBSclinical study.

For the Cyprus cohort, a follow-up was conducted until 2017, averaging 11 ± 3 years, during which cardiovascular events or deaths were recorded. All ultrasound scans from Cyprus were performed using a Philips (ATL) HDI 5000 duplex scanner, equipped with a broadband L12-5 MHz linear array transducer. Multiple images were taken for each subject, and the highest quality images, as determined by the operator, were selected for this study. The ultrasound parameters were standardized (central TGC, 170 dB dynamic range, low persistence, high frame rate) and remained consistent for all scans. In contrast, the Pisa participants' scans were conducted by a trained operator using a MyLab25 device (ESAOTE, Florence, Italy) with a LA523 4-13 MHz linear array transducer. For each side of the neck, one anterior and one lateral scan were performed, each lasting 10 seconds, and the best frame from these scans was chosen for analysis. A specific preset (central TGC, average 96 dB dynamic range varying from 84 dB to 102 dB, no persistence) was created and utilized for all examinations. In both protocols, the criterion for selecting the best images or frames was based on achieving the highest qualitative visual contrast between the carotid lumen and the intima-media complex. The pixel dimensions of the images in the dataset varied, with the most common being 0.064 mm/pixel, ranging from a minimum of 0.038 mm/pixel to a maximum of 0.267 mm/pixel. This pixel dimension, also known as the calibration factor (CF), is documented for each image in the publicly available dataset. Annotations for manual reference tracing were carried out on each image by an experienced analyst, A1 (L.Gi. from Torino, with over 10 years of experience in carotid sonography), and these annotations were regarded as the gold standard. The contours of the LI and MA anatomical interfaces were traced manually. A1 repeated the entire procedure 1 month later to evaluate intra-analyst variability (noted as A1s), and it was also performed by two additional expert analysts, A2 (G.V., with over 25 years of experience in carotid sonography) and A3 (for the Cyprus database: M.G., with over 25 years of experience in carotid sonography; for the Pisa database: L.Gh., with over 25 years of experience in carotid sonography), to measure inter-analyst variability. The annotations made by A1, A1s, A2, and A3 were kept blind from each other.

Center	Images	Conversion factor	Ultrasound equipment	Settings
Munich	100	0.073 mm/pixel	EPIQ 7G (Philips) L12-3 12-3 MHz linear array	Central TGC Gain 44-48 dB Dynamic range: 56 dB No persistence
Pisa	100	0.076 mm/pixel	MyLab25 (Esaote) LA523 4-13 MHz linear array	Central TGC Average 96 dB dynamic range No persistence No filters
Porto	100	0.062 mm/pixel	ATL HDI500 (ATL Ultrasound) L12-5 5-12 MHz linear array	Varying TGC and dynamic range No persistence No filters
Torino	100	0.064 mm/pixel	ATL HDI500 (ATL Ultrasound) L12-5 5-12 MHz linear array	Varying TGC and dynamic range No persistence No filters
Toronto (FAMUS)	100	0.045 mm/pixel	Linear array with central frequency = 7MHz*	No TGC Dynamic range: 45 dB No persistence No filters

Table 2 Acquisition parameters CUBS technical.

3.2.2. CUBStechnical

The second dataset, $CUBS_{technical}$, was an extension of the previous work, focused on comparing different models for automatic IMT measurement to understand which image feature could potentially decrease the performance of computerized methods [124].

For this study 500 images were collected. Specifically, 400 of these images were obtained from four distinct centers, with each center contributing 100 images using various ultrasound device systems and probes. The distribution of these images, along with the specifics of the systems and settings used at each center, is detailed in Table 1. Additionally, Figure 4 provides a representative image from each center for illustrative purposes. The Ethics Committee of each respective center granted approval for this study, and all participants involved provided their written informed consent. Acquisition protocol for all the participants followed the protocol discussed for the previous dataset.

In addition to the images obtained from clinical settings, the study also incorporated 100 simulated images. These were generated using the Fast And Mechanistic Ultrasound Simulation (FAMUS) software [125], [126]. For simulation, in-silico phantoms were created using a binary mask, with the IMC manually delineated along the width of the image. The FAMUS then simulated the final B-mode images, providing accurate ground truth profiles used for comparative analysis of both manual and computerized profiles. The simulated phantoms had dimensions of 40×1×25 mm^3 in the x, y, and z (lateral, elevation, and axial) directions, and incorporated 30,000 scatterers, equating to a density of 30 scatterers/mm^3. The employed simulated array transducer had a central frequency of 7 MHz with 192 active elements, and the B-mode image reconstruction was executed using 128 scanlines. Regarding the pixel dimensions of the images in the entire dataset, the mode was found to be 0.060 mm/pixel. The range of these dimensions varied, with a minimum of 0.029 mm/pixel and a maximum of 0.099 mm/pixel. The same procedure for the reference measurements was performed also for this dataset.

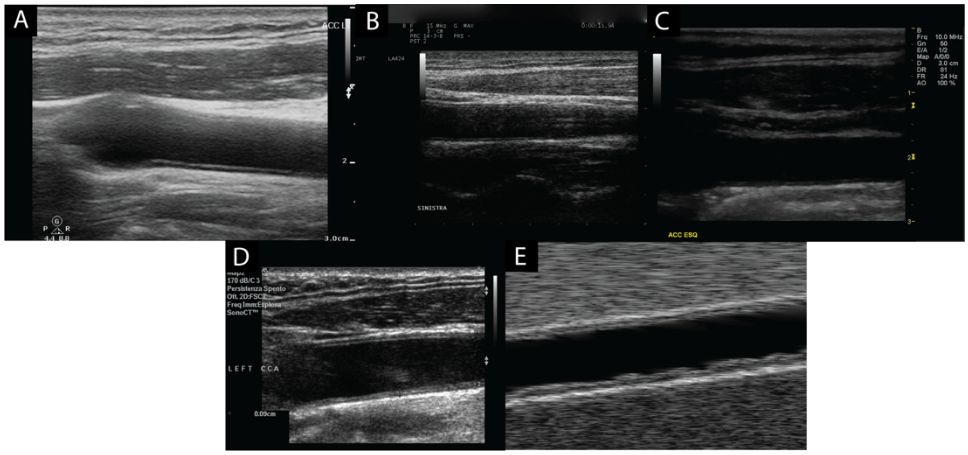


Figure 4 Example of images from each center (A-D) in-vivo images. (E) FAMUS simulated images. (A) Munich, Germany; (B) Pisa, Italy; (C) Porto, Portugal; (D) Torino, Italy; and (E) Toronto, Canada.

3.2.3. Multicenter

The third dataset was presented in [127]. In this study, ultrasound images were acquired from five distinct geographical and ethnic populations at five different institutions using four different original equipment manufacturer (OEM) ultrasound scanners. The demographics and acquisition parameters of this database are concisely summarized in Table 2. Prior to undergoing the ultrasound examination, all patients provided informed consent. Each institution was responsible for securing approval for the data acquisition from their respective Ethical Committee or Institutional Review Board (IRB).The datasets from the five institutions were acquired independently and in different years, as indicated in the first row of Table 3. Due to this independent acquisition process, there was no standardization among the five image sets. Moreover, the sonographers at each institution adjusted the scanner settings, including TGC values and gain factors, individually for each patient during the image acquisition process. All images were digitized at 8 bits. Consequently, the database is characterized as multi-ethnic, multi-institutional, and multi-OEM-scanner. It includes both healthy and pathological arteries, with images acquired at various positions along the carotid artery and customized to diverse gain settings. The manual segmentation of all images was performed by three expert vascular radiologists using ImgTracer™ software (Global Biomedical Technologies, Inc., CA, USA) [128]. Of this dataset 465 images from 4 different centers were selected because they were the ones for which three manual tracings were available hence a more robust estimation of IMT was possible.

	Torino (2002–2009)	**Nicosia (2002)**	**Cagliari (2009)**	**Porto (2008)**	**Hong Kong (2010–2011)**
Images	200	100	42	23	300
CF (mm/px)	0.0625	0.0600	0.0789	0.0900	0.0585
Patients	150	100	21	23	50
Age (years)	69 ± 16 (50–83)	54 ± 24 (25–95)	68 ± 8 (59–81)	[Not pub.]	60 ± 5 (54–67)
Scanner	ATLHDI5000	ATLHDI3000	Esaote MyLab 70	ATLHDI5000	Siemens Antares

Table 3 Acquisition parameters Multicenter.

3.3. IMT as a biomarker – clinical study

In the context of [105] the value of IMT as a biomarker was investigated. Five computerized CIMT algorithms were compared and assessed against manual measurements by three expert analysts on $CUBS_{clinical}$ dataset. The methods are based on the first-order absolute moment CNR_{IT} [129], anisotropic Gaussian derivative filters $INESCTEC_{PT}$ [109], dynamic programming TUM_{DE} [110], snakes UCY_{CY} [107] and dual snakes $POLITO_{IT}$ [130]. Inter- and intra-observer variability was assessed, and the values obtained by the computerized CIMT were compared with those obtained manually. The correlation of CIMT measurements with clinical parameters, cardiovascular event prediction through a generalized linear model, and the Kaplan–Meier hazard ratio was assessed using the measurements. The evaluation of IMT as a biomarker lead to mixed results depending on the method used to measure the biomarker. Figure 5 presents Kaplan–Meier plots for cardiovascular event risk analysis based on CIMT measurements. An increase in cardiovascular event risk is associated with higher CIMT groups, as evidenced by a statistically significant difference ($p < 0.05$) across most methods. However, this trend does not hold for the computerized technique UCY_{CY}, which appears to be an exception. Focusing on the method TUM_{DE} and on the one developed by our group ($POLITO_{IT}$), the hazard ratio analysis, detailed in Table 4, further describes these findings. This analysis, conducted by categorizing CIMT measurements into quartiles, reveals that the highest CIMT quartile is linked with a significantly elevated hazard ratio compared to the lowest quartile for almost all methods. Notably, the manual measurement method A3 did not follow this trend, suggesting that these methods might differ in how they correlate CIMT measurements with cardiovascular risk. The manual measurements A1 and A1', along with the computerized techniques TUM_{DE}, showed statistically significant higher hazard ratios for all three higher quartiles when compared to the lowest CIMT quartile. This consistency across quartiles indicates a strong correlation between increased CIMT measurements and higher cardiovascular event risk as assessed by these methods. These observations suggest that while most measurement methods correlate higher CIMT with increased cardiovascular risk, there are notable differences in how various techniques relate CIMT measurements to cardiovascular event risk. This discrepancy underscores the importance of method selection in clinical assessments of CIMT and its relationship to cardiovascular health.

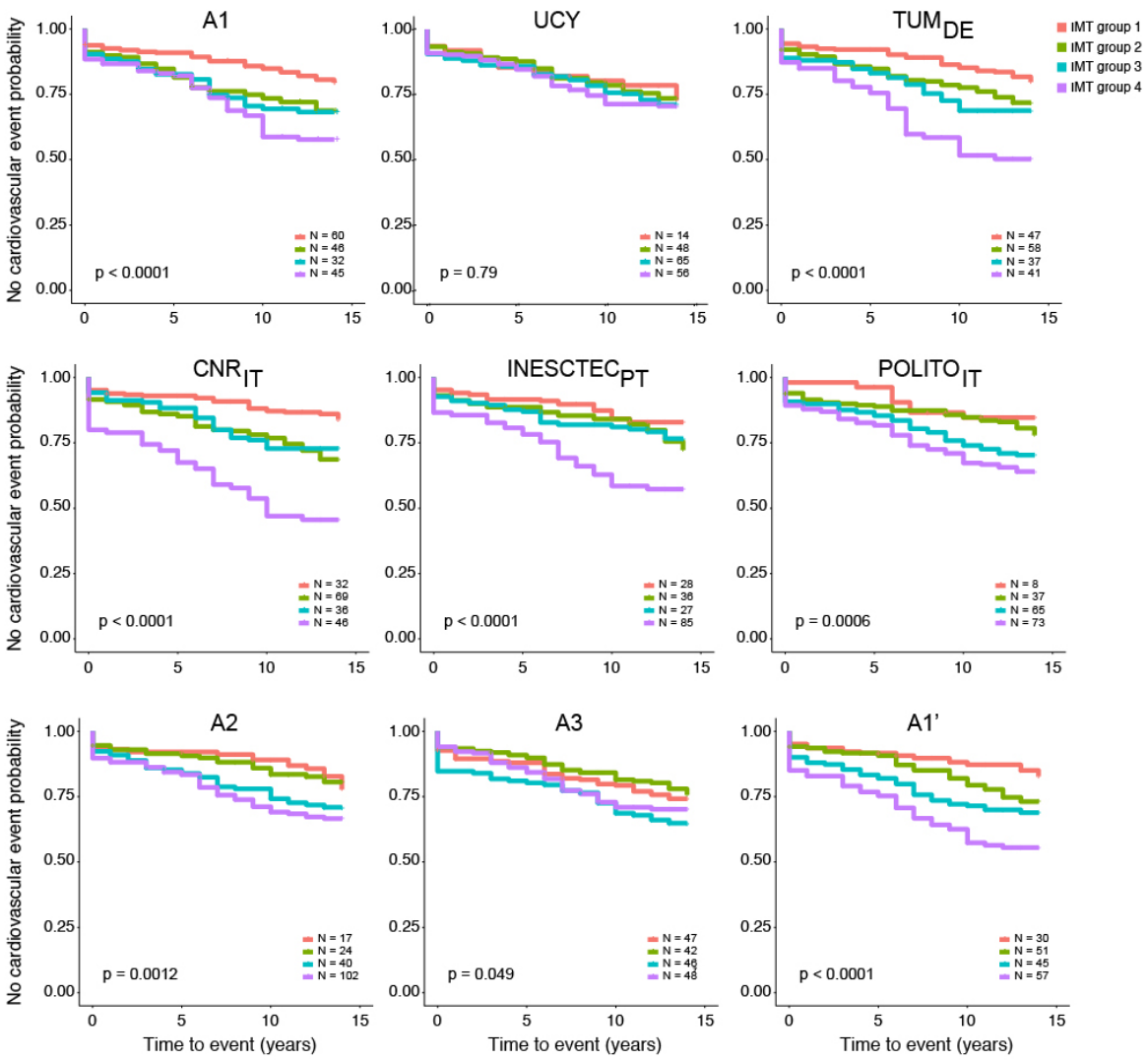


Figure 5 Kaplan–Meier cardiovascular event analysis plots, considering four CIMT groups: <650 µm, between 650 and 750 µm, between 750 and 850 µm and >850 µm. A1, A2 and A3 = manual segmentations of analysts 1, 2 and 3, respectively. A1' = segmentations of analyst 1, traced 1 month after A1; UCYCY = computerized method from University of Cyprus; TUMDE = method from Technische Universität München; CNRIT = method from Consiglio Nazionale delle Ricerche; INESCTECPT = method from INESC Technology and Science; POLITOIT = method from Politecnico di Torino.

Hazard group		A1	TUM_{DE}	$POLITO_{IT}$	A1'	A2	A3
IMT < 650 um	% Events	0.33	0.26	0.04	0.16	0.09	0.26
	H ratio	1	1	1	1	1	1
	95% CI	N/A	N/A	N/A	N/A	N/A	N/A
	p-value	N/A	N/A	N/A	N/A	N/A	N/A
	Bias (µm)	0±0	78±64	187±94	108±87	217±162	185±151
IMT ≥ 650 um IMT < 750 um	% Events	0.25	0.32	0.20	0.28	0.13	0.23
	H ratio	1.7	1.6	1.3	1.8	1.1	0.84
	95% CI	1.2-2.5	1.1-2.4	0.6-2.8	1.1-2.8	0.6-2.1	0.55-1.3
	p-value	0.006	0.014	0.514	0.010	0.736	0.402
	Bias (µm)	0±0	70±58	123±91	92±71	202±170	143±136
IMT ≥ 750 um IMT < 850 um	% Events	0.17	0.20	0.36	0.25	0.22	0.25
	H ratio	1.8	1.9	2.1	2.3	1.8	1.47
	95% CI	1.2-2.8	1.2-2.9	1.0-4.4	1.5-3.7	1.0-3.2	0.98-2.2
	p-value	0.005	0.003	0.490	<0.001	0.039	0.063
	Bias (µm)	0±0	71±52	91±71	105±79	169±131	149±105
IMT ≥ 850 um	% Events	0.25	0.22	0.40	0.31	0.56	0.26
	H ratio	2.5	3.4	2.6	3.7	2.2	1.21
	95% CI	1.7-3.7	2.2-5.2	1.3-5.5	2.4-5.7	1.3-3.7	0.81-1.8
	p-value	<0.001	<0.001	0.009	<0.001	0.003	0.346
	Bias (µm)	0±0	173±402	166±387	170±386	214±396	268±398

Table 4 Hazard ratio results for cardiovascular events (N = 183). A1, A2, A3: manual segmentations of analysts 1, 2, and 3, respectively. A1': segmentations of analyst 1, traced one month after A1. TUMDE: method from Technische Universität München; POLITOIT: method from Politecnico di Torino. H ratio: Hazard ratio. CI: confidence interval. The bias is expressed in absolute value.

3.4. Hybrid Ground Truth

The training target for networks, commonly known as ground truth, is usually based on the manual segmentation performed by a single operator. This method inherently skews the network predictions towards the preferences of that specific operator. Before deep learning, the field of automatic IMT measurement was extensively explored. These pre-deep learning studies, which primarily utilized heuristic-based models, should not be overlooked, since they have potential value in improving the performance and generalizability of deep learning systems.

In our research, an innovative hybrid learning approach for IMT measurement was explored. This method effectively integrates the strengths of both heuristic and deep learning methodologies. It does so by merging manual segmentations with the robust automatic segmentation achieved through a

dynamic programming technique, as detailed in the reference [110]. The segmentation obtained through this combined approach is then used as the ground truth for training and validating the model.

The hybrid ground truth concept involves a profile consensus strategy. This strategy aims to establish a collectively determined hybrid ground truth (Hybrid GT) profile. The process involves a separate column-wise consensus for the LI and MA borders. To assess the compatibility between two profiles (designated as P1 and P2), a similarity parameter is utilized. This parameter quantifies consistency based on their shared columns – the image columns where both profiles are defined. The mathematical formulation of this concept is presented in Equation (1).

$$Similarity\ (P_1; P_2)\ =\ \frac{Correlation\ (P_1; P_2)}{Bias\ (P_1; P_2)} \tag{1}$$

The calculation of $Correlation\ (P_1; P_2)$ is based on the row values of the LI/MA profiles. Conversely, $Bias\ (P_1; P_2)$ represents the average column-wise deviation between the two profiles. The observed similarity range was [0, 3.30] for the LI profile and [0, 3.82] for the MA profile. Profiles were classified as similar if the calculated Similarity exceeded the empirically determined threshold of 0.5. Considering two manual segmentations (A1 and A1s) and one computerized segmentation (TUM), the Similarity parameter was evaluated for each of the three pairings: A1 vs. A1s, A1 vs. TUM, and A1s vs. TUM.

In determining the Hybrid GT profile, five potential scenarios are outlined in Figure 6:

- Absence of both manual and computerized profile points, leading to an undefined Hybrid GT profile point.
- Presence of a single profile, which then defines the Hybrid GT.
- Presence of two manual profiles, where the Hybrid GT is the average of the two if they are highly similar. If not, the profile most similar to the computerized method is selected.
- Presence of one manual and one computerized profile. If they are highly similar, the Hybrid GT comprises both; otherwise, only the manual profile is retained.
- Presence of all three profiles, where the Hybrid GT point is the average of the two profiles showing the highest similarity among the three.

Figure 6 provides a visual representation of the process, highlighting the decision-making differences between the LI and MA profiles. In this figure, the LI profiles from two manual sources are deemed similar and thus combined to form the Hybrid GT. Conversely, the two manual MA profiles display low similarity. Therefore, the Hybrid GT for the MA is derived by averaging the A1 profile with the TUM profile, as it showed the greatest similarity with the TUM.

The deep learning system is comprised of two main components. The first component is designed to identify the ROI that encompasses the IMC using a detection network. The second component employs a patch-based network for segmenting the IMC, as depicted in Figure 7. To evaluate the effectiveness of the proposed method, the Multicenter dataset was utilized. This dataset presents a significant challenge in terms of generalization for the trained network. For the training of both the detection and segmentation models, the first two datasets described earlier were used. These datasets were divided in a 90/10 split, resulting in 2311 images for training and 265 for validation. Initially, each original image was cropped using a heuristic approach to focus solely on the relevant ultrasound section. Subsequently, the cropped image was resized to a resolution of 480x480 pixels. The same processing steps were applied to the mask generated using the Hybrid GT profile.

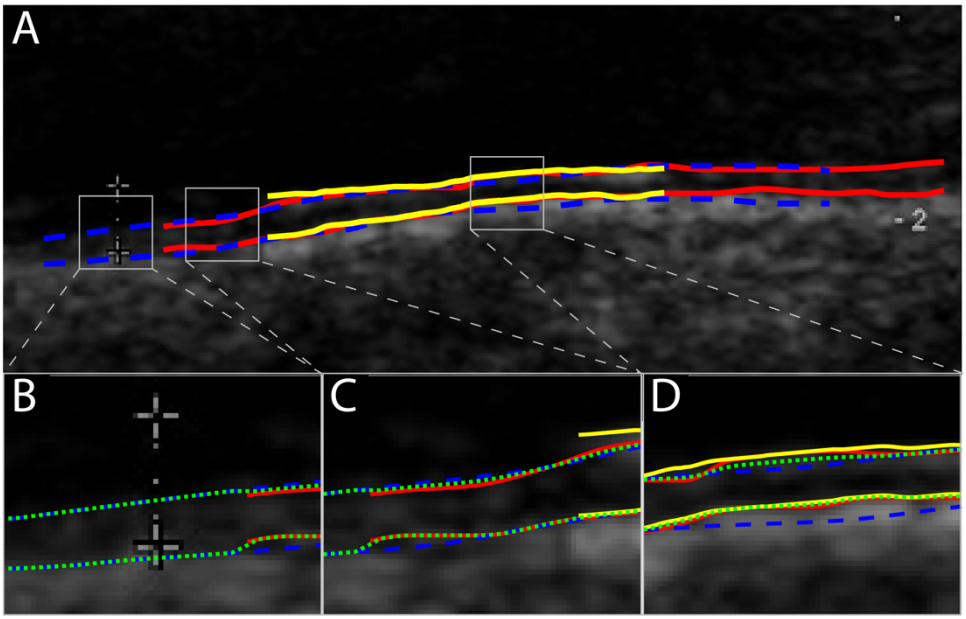


Figure 6 Illustration of the Hybrid GT construction. The profiles are A1 (red), A1s (blue dashed), TUM (yellow), Hybrid GT (green dotted). In panel A there are the three available profiles (two manual, A1 and A1s, and one computerized, TUM). In panel B, the condition where only one profile is available is shown. In panel C, two profiles are available. In panel D, three profiles are again available.

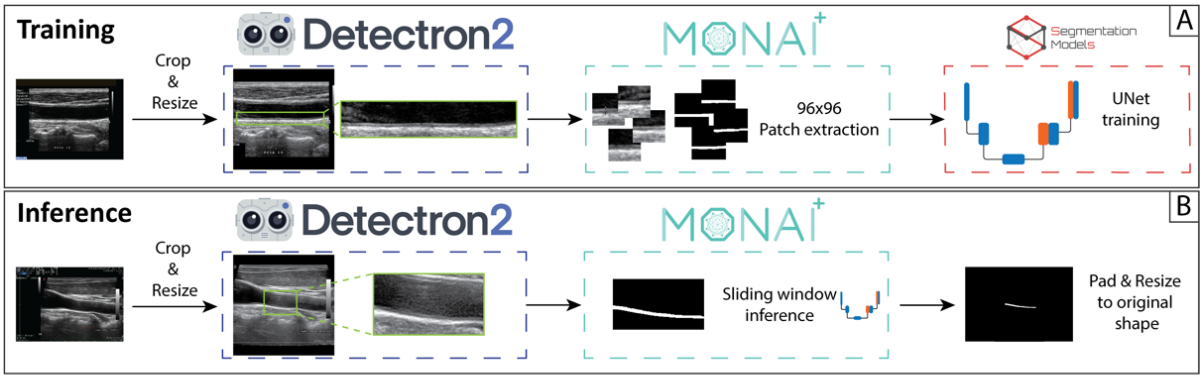


Figure 7 Training (top row) pipeline and inference (bottom row) pipeline.

A Faster R-CNN model, referenced in [131] was implemented as the detection network. This network is proficient in predicting object boundaries and assigning scores to each detected bounding box. The Faster R-CNN consists of two integral parts: the first is a deep convolutional network (ResNet50) that proposes regions, while the second is a detector [132] that further analyzes and classifies these proposed regions. The detection network utilizes anchor boxes, which are reference boxes of specific scales and aspect ratios. In our case, we used five different reference anchor boxes ([150, 170, 190, 220, 240]) with three aspect ratios ([2, 3.5, 5]) to efficiently detect objects of varying shapes. The network underwent training for 20 epochs, with an early stopping mechanism activated after 5 epochs of stagnant performance on the validation set. The learning rate was set at 10^-3, and the batch size was fixed at 6. Following this, up to six 96x96 patches were extracted from the output of the detection network using MONAI (Medical Open Network for AI), as cited in [133]. This extraction resulted in 13862 patches for the training set and 1590 for the validation set, which were then used for training the segmentation network. For the segmentation phase, we utilized a standard UNet framework with a pretrained Resnet50 encoder. The training parameters included a batch size of 32, Adam optimizer, a learning rate of 0.0002, and Dice loss for optimization. To enhance the model's robustness and

	$UNet_{A1}$	$Unet_{A1s}$	$Unet_{TUM}$	$Unet_{HybridGT}$
DICE	0.893 ± 0.043	0.877 ± 0.045	0.897 ± 0.051	**0.907 ± 0.037***
HD mask (pixel)	3.53 ± 2.53	3.98 ± 2.23	3.75 ± 2.36	**3.21 ± 2.14***
ABS IMT bias (mm)	0.153 ± 0.088	0.195 ± 0.107	**0.053 ± 0.088***	0.136 ± 0.079
HD LI (mm)	0.176 ± 0.105	0.199 ± 0.140	0.199 ± 0.211	**0.158 ± 0.092***
HD MA (mm)	0.301 ± 0.238	0.327 ± 0.226	**0.232 ± 0.208***	0.272 ± 0.204

*Table 5 Results summary – common support. HD mask: Hausdorff distance between the reference mask and the predicted mask; ABS IMT bias: absolute intima-media thickness bias; HD LI(MA): Hausdorff distance between the reference LI(MA) profile and the predicted LI(MA) profile. UNetA1, UNetA1s, UNetTUM, UNetHybridGT: UNet models trained with the specified ground truth as the target. In bold the best performing method and the * indicates a significant difference with the other methods (Wilcoxon Test, p <= 0.01).*

versatility, we applied online data augmentation on the training patches. This augmentation involved a random combination of various transformations: Scaling and Rotation (scale changes within ±10%, rotation within ±10°, probability = 1), Horizontal Flipping (probability = 0.5), Gaussian Noise Addition (probability = 0.3), Multiplicative Noise Addition (probability = 0.2), Contrast Limited Adaptive Histogram Equalization (CLAHE) (probability = 0.3), Sharpening (probability = 0.25), Optical Distortion (probability = 0.2), and Blurring (probability = 0.5). The optimal model was selected based on its performance on the Intersection over Union metric on the validation set. This method was developed using the Pytorch framework, in conjunction with the openmmlab, Albumentations, and Segmentation Models libraries [134], [135], [136]. The training was conducted on an RTX 3090 with 24 GB of VRAM, taking 1.5 hours for the detection model and 0.5 hours for the segmentation model.

During the inference phase on the external test set, the same sequence of detection, patch extraction, and segmentation was employed. The network's prediction was then scaled and upsampled to the original dimensions of the image for extracting the LI and MA profiles, aligning with the manual profile reference system. The largest connected area's boundary was identified, and its perimeter coordinates were smoothed using a moving average filter with a 5-pixel window. To evaluate which learning approach (manual, computerized, or Hybrid GT profile) yielded the best results on the test set, the same network was trained with four different targets (A1, A1s, TUM, and Hybrid GT). The UNet's predictions were labeled according to their respective training targets (UNetA1, UNetA1s, UNetTUM, UNetHybridGT).

The test set's reference mask and LI and MA profiles were derived from the consensus of three independent manual tracings (GT1, GT2, GT3). This consensus was established using the methodology described earlier for creating the hybrid ground truth. The performance of the segmentation framework was evaluated using the Dice coefficient and the Hausdorff distance (HD) between the predicted and reference masks. The precision of LI and MA boundary segmentation was specifically assessed by measuring the HD between the predicted and reference profiles. Additionally, the absolute IMT bias was calculated using the following equation:

$$Abs.\ IMT\ Bias\ = |IMT_{method} - IMT_{Ground\ truth}| \quad (2)$$

In this equation, 'method' refers to the automatic segmentation, and 'Ground truth' is the IMT measured from the consensus of the three manual operators. The validation metrics, computed based on the common support (columns where both the target and all predicted profiles are defined), are presented in Table 5. To determine the statistical significance of the differences between the best-performing method and others, a Wilcoxon test was conducted, and the non-normality of the distribution was verified using a chi2 test. To further validate the generalization performance on the test set, the IMT measurements from the manual operators with those from the hybrid learning

paradigm were compared. The intraclass correlation coefficient (ICC(2,1)) was calculated in two scenarios: (1) between the IMT values from the three manual operators, and (2) between the IMT value from UNetHybridGT and the average IMT value of the three operators. To compare the UNetHybridGT's agreement with manual operators against inter-operator variability, a Bland-Altman analysis was performed. The Hybrid GT method we introduced demonstrated superior segmentation performance, evidenced by a Dice score of 0.902 ± 0.047 and a HD of 3.53 ± 3.05 mm between segmentation masks. In the context of tracing the LI profile, the UNetHybridGT achieved better results, with a Hausdorff Distance of 0.158 mm ± 0.092 mm. However, it's noteworthy that for specific metrics, the network trained solely with the TUM ground truth exhibited more favorable outcomes. In terms of the absolute bias in IMT, the TUM-based network recorded a lower bias of 0.053 mm ± 0.088 mm compared to the 0.136 mm ± 0.079 mm of the UNetHybridGT. Furthermore, for the tracing of the MA profile, the TUM-trained network also surpassed the novel approach, registering a Hausdorff Distance of 0.232 mm ± 0.208 mm against the 0.272 mm ± 0.204 mm of the UNetHybridGT. This demonstrates that while the Hybrid GT is generally effective, certain aspects, such as the IMT bias and MA profile tracing, could be better addressed with the TUM ground truth.

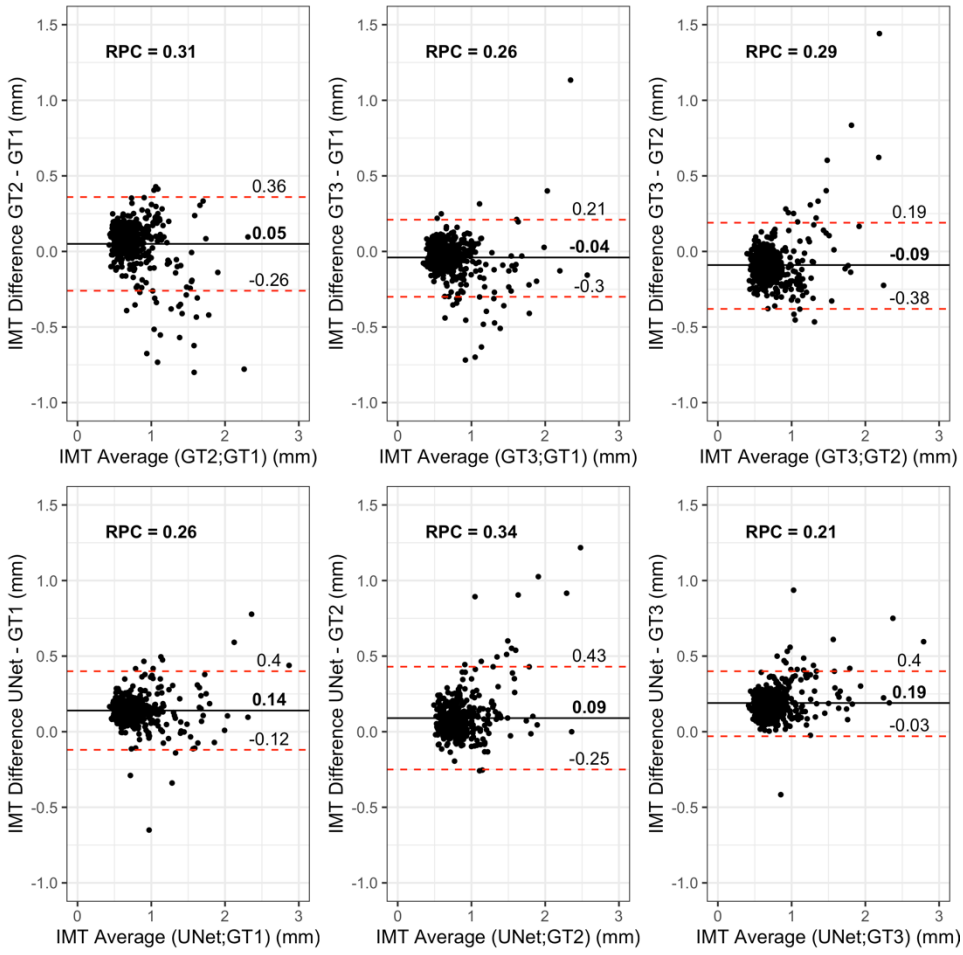


*Figure 8 Bland-Altman plots comparing inter-operator variability and UNetHybridGT results on the test set. Manual measurements are referenced as GT1, GT2, GT3. UNetHybridGT is referenced as UNet. RPC: reproducibility coefficient (1.96*standard deviation of the errors). The horizontal lines define the mean error (black solid line), and upper and lower limits (dashed red lines, ± 1.96 SD). In bold the mean differences with a significant statistical difference (p<0.01).*

The robustness of our new approach was validated through a correlation analysis comparing manual and automatic IMT measurements. The Pearson's correlation coefficients for IMT values between the UNetHybridGT and the three manual operators (GT1, GT2, and GT3) were highly positive: 0.944 for GT1, 0.889 for GT2, and 0.944 for GT3. These results indicate a strong linear relationship between the automated measurements and each of the manual tracings. In contrast, the correlations between the manual operators themselves were slightly lower: 0.863 for GT1 vs. GT2, 0.925 for GT1 vs. GT3, and 0.827 for GT2 vs. GT3. These values, while still indicative of good agreement, are not as high as those achieved by the UNetHybridGT in comparison with the manual operators. Furthermore, the agreement between the UNetHybridGT's measurements and the average of the manual measurements was assessed using the ICC(2,1). This analysis yielded an ICC of 0.958, demonstrating excellent agreement. This level of agreement significantly surpasses that observed among the three manual operators, which had an ICC of 0.859.

A Bland-Altman analysis, illustrated in Figure 8, revealed a slight overestimation of the IMT value by the UNetHybridGT. However, no significant bias was observed in the distribution of errors. The reproducibility coefficient (RPC) was comparable between the manual and automatic methods, suggesting that the automatic approach's performance falls within the range of variability typically seen between different human operators. This finding further supports the reliability and accuracy of the UNetHybridGT in replicating manual IMT measurement methods.

Similar results were obtained in [137] where the TUM ground truth was compared directly with manual ground truth. The segmentation was performed directly on the whole image and TUM resulted as the best target when testing network generalization.

3.5. Common Dice Loss

The growing interest in advanced deep-learning architectures with hybrid mechanisms for diverse applications has been notable in recent years. Despite this, the development of innovative loss functions, which could further enhance model performance, has not received equivalent attention. Traditional loss functions in deep learning models quantify the discrepancy between the predicted output and the true target, a process essential for optimizing model parameters during training. However, these conventional loss functions come with several limitations. These include the unchanging weights of individual losses throughout training, the focus on a single objective that may not fully represent complex input-output relationships, and the uniform treatment of all image parts regardless of task difficulty. This equal treatment of all image segments can lead to suboptimal results, particularly in scenarios where specific regions require more focus due to their complexity in segmentation.

Recent research, especially in the realm of ultrasound imaging, has introduced a variety of loss functions that address these shortcomings. These novel functions broaden the scope of the optimization challenge beyond region-based losses like Dice. They encompass various segmentation priors such as shape [138], structure [139], uncertainty [140], class-imbalance [141], and boundary awareness [142], [143].

The aim of this section is to evaluate the effectiveness of a loss function in segmenting fine structures in US images. A common-support Dice loss function is proposed. The underlying hypothesis is that a common-support Dice loss function could more accurately address the challenges inherent in US image segmentation, such as pixel imbalance and precise edge detection.

Two open-access datasets, $CUBS_{clinical}$ and $CUBS_{technical}$, were employed for training the network. The dataset was randomly divided in an 80/20 ratio to create training and validation sets. The semi-

automatic segmentation method's output TUM was used as ground truth during the training phase, as it was found to be the most effective in [137] with a whole image approach. For evaluation, the Multicenter dataset was utilized to test the approach. The segmentation task was framed as a binary classification problem, distinguishing between "Background" (BG) and "Intima-Media-Complex" (IMC), using a UPerNet [144] with a ConvNeXt-Small encoder [145]. The Adam optimizer was set with an initial learning rate of 10^-5 and a batch size of 4 samples for the training process. Sample augmentation was done using geometric and intensity transformations. The network underwent 11400 iterations over 20 epochs, with performance metrics logged each epoch to select the best-performing model.

The primary objective of this research was to identify the most effective loss function for this specific task. Consequently, the network was trained with various objective functions, including Dice loss, Focal loss [146], and Lovasz loss [145], to benchmark the performance of our proposed loss functions:

Generalized Dice Loss (GDL): This is a modified version of the loss function presented in [147], originally developed to address highly unbalanced segmentation problems in 2D and 3D medical images. Unlike the original implementation, this version recalculates the weighting of the two classes (w_k) for each batch. The total number of pixels in each batch is denoted by N, with i representing the i-th pixel. C represents the number of classes, with k indicating the k-th class. T is the target segmentation, and P represents the network prediction post-SoftMax operation. The equation is as follows:

$$L_{gdl} = 1 - 2\frac{\sum_{k=1}^{C} w_k \sum_{i=1}^{N} T_i^k P_i^k}{\sum_{k=1}^{C} w_k \sum_{i=1}^{N} (T_i^k)^2 + (P_i^k)^2} \tag{3}$$

$$w_k = \frac{1}{(\sum_{i=1}^{N} T_i^k)^2} \tag{4}$$

Common-support Dice Loss (CDL): This function modifies the previous one by applying a mask (M) to both the target and predicted tensor. The mask is positive for pixels in columns where both tensors have at least one pixel labeled as IMC. This approach focuses the loss calculation on image areas containing the most information. The masking process also avoids penalizing the network for predicting a wider segmentation area than the ground truth, considering the semi-automatic annotation's conservative nature.

$$L_{cdl} = 1 - 2\frac{\sum_{k=1}^{C} w_k \sum_{i=1}^{N} M \times T_i^k M \times P_i^k}{\sum_{k=1}^{C} w_k \sum_{i=1}^{N} (M \times T_i^k)^2 + (M \times P_i^k)^2} \tag{5}$$

Common-support Boundary Loss (BL): Adapted from the loss first presented in [148], this loss uses the same common-support method as the CDL. Here, D_i^k and P_i^k result from the distance transform of the ground truth and predicted segmentation, respectively.

$$L_{bl} = \frac{1}{NC} \sum_{k=1}^{C} \sum_{i=1}^{N} M \times D_i^k M \times P_i^k \tag{6}$$

The evaluation of these loss functions aims to enhance the accuracy and precision of US image segmentation, particularly for fine structures like the IMC, which are crucial for clinical diagnosis and

Loss	Dice	HD	HD LI	HD MA	IMT Abs. Bias
Dice	0.900 ± 0.057	0.210 ± 0.510	2.804 ± 6.686	3.006 ± 6.497	0.811 ± 0.842
Focal	0.899 ± 0.040	0.184 ± 0.101	2.292 ± 1.723	2.488 ± 1.591	0.830 ± 0.841
Lovasz	0.899 ± 0.040	0.187 ± 0.102	2.452 ± 1.850	2.718 ± 1.839	0.840 ± 0.889
GDL	0.900 ± 0.040	0.189 ± 0.104	2.478 ± 1.969	2.621 ± 1.913	0.843 ± 0.882
CDL	**0.903 ± 0.036**	**0.182 ± 0.097**	**2.136 ± 1.348**	**2.342 ± 1.415**	**0.805 ± 0.901**

treatment planning. To validate the performance of the proposed approach, we adopted the same approach used in the previous section. For segmentation performance specifically, Dice score and

Table 6 Metrics calculated on the Test set for the same network trained with different objective functions. GDL: Generalized Dice Loss. CDL:Common-support Loss. Hausdorff Distances (HD) and Intima Media Thickness (IMT) are expressed in pixels.

Hausdorff Distance were calculated on the external test set. After extracting the LI and MA profile coordinates and measuring the IMT, for each image the IMT absolute bias was calculated as well as the Hausdorff Distance for both the LI and MA profiles separately. These metrics were computed on the common support between all the methods in the comparison. Results are presented in Table 6, distance-based metrics are expressed in pixels. Results of the network trained only with the common support boundary loss are not shown since the network couldn't reach a convergence point.

The network trained with the CDL showed improvements in all the metrics taken into consideration. These improvements are more conclusive in distance-based metrics where the Hausdorff Distance for LI and MA profile is more than 15% lower on average with respect to the second-best performing loss function.

3.6. Discussion

Previous studies have emphasized the criticality of assessing Common Carotid Artery (CCA) segmentation in image regions where both manual and automatic methods coincide [124]. This approach is vital because it targets the most clinically relevant regions, ensuring that the automated methods are benchmarked against the highest quality segments of manual tracings. This common ground provides a more realistic and challenging test for automated segmentation techniques, particularly in the context of ultrasound images where variability and noise are prevalent. While manual segmentations by expert radiologists are considered the gold standard in medical imaging, they are not devoid of limitations. Variability exists not only between different operators but also within the same operator over time. Such variability, attributed to human error or subjective interpretation, impacts the accuracy and reliability of these segmentations. This variation becomes especially significant when these manual segmentations are used for training and validating deep learning networks, as any inconsistency can propagate through the model, affecting its performance and generalizability. Given the inherent variability in manual segmentations, there is a pressing need for a more robust ground truth. This robustness can be achieved by integrating manual tracings with automated, algorithm-based approaches, leading to a more stable and consistent ground truth for model training and validation. Additionally, validating these models against segmentations from multiple operators is crucial for assessing the generalization and reliability of the models. This multi-operator approach helps in averaging out individual biases and errors, providing a more comprehensive evaluation of the model's performance. The introduction of Hybrid Ground Truth (HybridGT) marks a significant advancement in creating a more stable and reliable ground truth for deep learning models. By combining the strengths of manual expertise and algorithmic precision, HybridGT addresses the limitations of both approaches. The evaluation of the Hausdorff Distance (HD)

between the profiles generated by UNetHybridGT and those obtained through the consensus of three manual profiles yielded mixed results. For the Lumen Intima (LI) profile, UNetHybridGT outperformed the others, demonstrating greater accuracy. However, in the tracing of the Media Adventitia (MA) boundary and in the measurement of the intima-media thickness (IMT), UNetHybridGT was less precise compared to $UNet_{TUM}$. Despite this, UNetHybridGT still showed better performance than both $UNet_{A1}$ and $UNet_{A1s}$ and exhibited a lower standard deviation in comparison to $UNet_{TUM}$, indicating its robustness.

Furthermore, in the test set, the mean conversion factor was 0.0618 ± 0.0109 mm/pixel. Thus, the difference in HD for the MA profile between $UNet_{TUM}$ (0.232 ± 0.208 mm) and UNetHybridGT (0.272 ± 0.204 mm) is less than one pixel, which is a relatively minor discrepancy.

One plausible explanation for these results is that the MA boundary in ultrasound images is typically less affected by noise and can often be traced accurately using heuristic-based computerized methods. Consequently, a network trained solely with this information may transfer knowledge more effectively to an external test set. On the other hand, tracing the LI profile is more reliant on human expertise, suggesting that networks trained with a mix of computerized and manual examples can delineate this boundary more precisely.

Incorporating information from a computerized method in the training examples allows the network to learn more defined rules, reducing errors in segmenting targets that are well-characterized by these rules, such as the MA boundary. This effect is also observed when training the network with a ground truth derived solely from the computerized method. In such scenarios, although the network may lose some generalization capacity, it achieves improved performance in capturing the finer details of the MA profile segmentation. This phenomenon underscores the importance of choosing appropriate training data and methods to balance generalization and precision in deep learning models, especially in complex tasks like ultrasound image segmentation.

The implementation of a common-support Dice loss (CDL) function is a novel approach to optimizing deep learning models for ultrasound image segmentation. This loss function, by focusing on the regions of the image where both manual and automated methods agree, directs the model's learning towards the most critical and challenging parts of the segmentation task. It ensures that the model does not get biased towards over-segmentation or under-segmentation, common pitfalls in medical image segmentation.

In conclusion, this chapter has demonstrated the importance of a robust ground truth and the effectiveness of specialized loss functions in enhancing the performance of deep learning models for CCA segmentation in ultrasound images.

4. Optic Nerve Characterization in Transorbital Ultrasound Images on a Multicenter Data Set

4.1. Introduction

Assessing the Optic Nerve Diameter (OND) and Optic Nerve Sheath Diameter (ONSD) through transorbital sonography (TOS) offers a valuable non-invasive method for detecting various neurologic conditions by estimating intracranial pressure (ICP)[149]. Although the gold standard for ICP monitoring is its invasive assessment, this method isn't suitable for emergencies or for centers lacking neurosurgical equipment.

The Optic Nerve (ON) is an extension of the dura mater and houses cerebrospinal fluid. Due to the ON sheath's elasticity, any ICP elevation leads to a proportional increase in the ONS, with a corresponding rise in the observed ONSD[150], [151]. Research [152] indicated that the ONSD values derived from B-mode US images correlated well with ICP values, showing a correlation coefficient (R) of 0.76 among patients in a neurocritical care unit. Notably, ONSD measurements have demonstrated their worth in diagnosing intracranial hypertension (ICH) when ICP values exceed 20 mmHg [153]. A significant benefit of TOS is its adaptability for use in bedside or critical care scenarios [149], [154], and its feasibility even with compact devices, widening its potential applications [155]. Additional OND and ONSD uses include monitoring multiple sclerosis [156], brain infections [157], and demyelinating ailments [158].

While manual OND and ONSD evaluations show consistent reliability, ultrasound image acquisition, and interpretation necessitate expertise [159], [160], [161]. Streamlining the image acquisition and analysis procedures in TOS could enhance the examination in several ways, such as optimizing machine settings and automating image processing, improving accuracy and efficiency.

Recent studies have introduced various automated or semi-automated image segmentation methodologies in TOS for OND and ONSD measurement. Gerber et al. [162] showcased a two-phase technique, leveraging a 3D printed phantom with a mobile US probe. Meanwhile, Soroushmehr et al. [163] employed a segmentation strategy based on superpixel clustering. Another approach was by Meiburger et al. [164], which utilized a Gaussian derivative filter-based method. Stevens et al. [165] incorporated active contours, utilizing a signed asymmetry map for initializations. Our team introduced a Deep Learning (DL) methodology for this purpose in 2021, revealing its superiority over traditional techniques by achieving a Mean Absolute Error of 0.48 ± 0.48 mm in ONSD assessments [166].

The following section presents the work [167] which had the purpose of further validating the efficacy of the DL model, testing it on diverse datasets from four separate machines. Motivated by the limited numerosity of the available images and the high variability of the dataset, in this research an ensemble of UNet networks has been explored to achieve acceptable segmentation performances. Contrary to previous research where the reference OND and ONSD were strictly from manual segmentation masks, this study compares automated measurements against both manual segmentation masks and manual diameter calculations via a dedicated graphical user interface (GUI). The impact of the Signal-to-Noise Ratio (SNR) on the segmentation and diameter evaluation effectiveness is also compared, aiming to pave the way for standardizing the acquisition process for training segmentation models.

4.2. Database Description

For this research, we utilized a proprietary dataset containing 464 TOS ultrasound images in .bmp format. These images were taken from 110 distinct subjects using four separate ultrasound machines. A subset of this database was previously utilized in our earlier research, in which the deep learning (DL) method was compared against conventional heuristic-driven techniques. Specifically, 207 of the images were acquired with the Esaote MyLab Gold 30 and MyLab Seven devices in Homburg Medical Centre; 113 were taken using the Esaote MyLab in Turin; 93 were sourced from the Toshiba Medical Systems Aplio 300 and AplioXG; and the remaining 51 images came from the GE-Vivid7 sonography system.

These ultrasound images were captured by expert sonographers using linear transducers that operated within central frequency values of 3-15 MHz. In terms of resolution, the Esaote devices in Turin and Homburg achieved lateral resolutions of 0.47 mm and 0.40 mm at frequencies of 10 MHz and 15 MHz, respectively. However, we didn't have similar resolution information for the Toshiba Aplio 300, AplioXG, and GE-Vivid7 systems. Across all the devices, the median pixel-to-distance conversion factor stood at 0.1061 mm/pixel, with a range between 0.0527 mm/pixel and 0.1210 mm/pixel.

To provide a clearer visual understanding, samples from each dataset are displayed in Figure 9. Additionally, detailed specifications of each device can be found in Table 7. It's essential to note that this research is retrospective. All the images were taken during standard clinical procedures and were subsequently de-identified to ensure privacy. The local institutional review boards of the participating centers did not mandate any additional approvals.

Device	Linear Probe frequency (MHz)	Lateral Resolution	Conversion factor (mm/pixel)	Numerosity
MyLab Esaote	10 MHz	0.6 mm (6.6 MHz) 0.47 mm (10 MHz)	0.1029 ± 0.0124 mm/pixel	207
MyLab Esaote	10 MHz	0.6 mm (6.6 MHz) 0.47 mm (10 MHz)	0.0848 ± 0.0120 mm/pixel	113
Toshiba Aplio300 and AplioXG	7.2-14 MHz	N/A	0.1020 ± 0.0108 mm/pixel	93
GE-Vivid7	7- to 11 MHz	N/A	0.1124 ± 0.0097 mm/pixel	51

Table 7 Summary of device characteristics. N/A (Not Available).

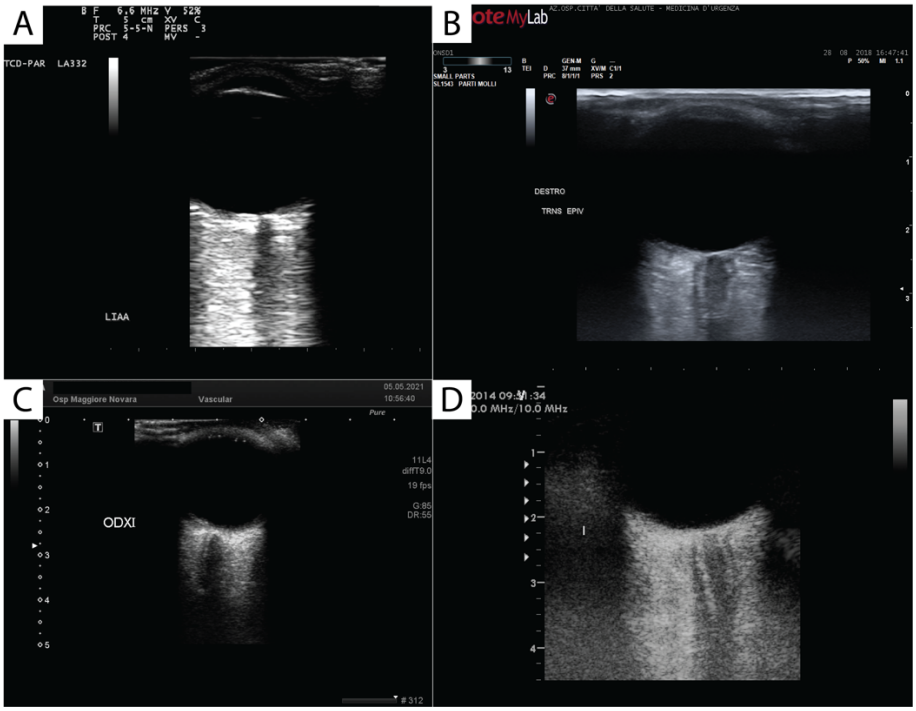


Figure 9 Examples of images available in our dataset from different centers and machines. A) Homburg, Germany (MyLab Gold 30). B) Turin, Italy (Esaote MyLab). C) Novara, Italy (Toshiba Medical System Aplio 300). D) GE-Vivid7.

4.3. Manual and Automated OND and ONSD Measurement

Before performing manual measurement, all ultrasound images were automatically cropped. A 256 x 256-pixel window was centered on the Optic Nerve, based on methods briefly covered in [166]. The cropping makes use of pixel intensity and a calibration factor (CF). The first cropping step aimed to remove ancillary information present in clinical ultrasound images and corresponded to 15 % and 90 % of the height of the image. Within this specific range, the row with the highest average intensity value was identified. Following this, to crop the image along its width, the indices for the columns were determined as the boundaries of the longest consecutive sequence of non-zero pixels in the identified row. This approach allows for a precise and adaptive cropping mechanism based on pixel intensities and continuity. The result of this action is shown in Figure 10a (green box). Subsequent processes identified the optic bulb using a pixel value threshold and the circular Hough transform (CHT) [168].

The automatically cropped images were then manually segmented using MATLAB's ImageLabeler tool. Three labels were used for differentiation: background, optic nerve sheaths, and the optic nerve. Manual labeling involved outlining the optic nerve sheaths where they were clearly distinguishable. The optic nerve label was then retrieved by creating a mask from the optic nerve sheath convex hull and subtracting them. Visualization can be seen in Figure 10 and Figure 11.

To establish a diameter measurement benchmark, a graphical user interface (GUI) was developed. This helped measure the OND and ONSD in a 3 mm zone, mirroring standard clinical practices [169]. A line representing the optic nerve's axis was generated, and using the image's conversion factor, the 3 mm measurement was established. The GUI then allowed the operator to mark two straight lines

for diameter measurements of the OND and the ONSD. Details on this method are further illustrated in Figure 11 and Figure 12.

Two expert operators, A1 (10+ y of experience in clinical practice of TOS) and A2 (with 5 y of experience in ultrasound image analysis, who trained on 25 independent examinations and performed the ONSD measurement in a standardized view of the ONSD), took manual diameter measurements. These measurements acted as a measure of inter-rater variability and verified the accuracy of A2's manual segmentations. The geometric approach used for manual measurement was mirrored for automated OND and ONSD measurements. Screenshots of this process can be seen in Figure 11. Data and manual measurements are accessible at [170].

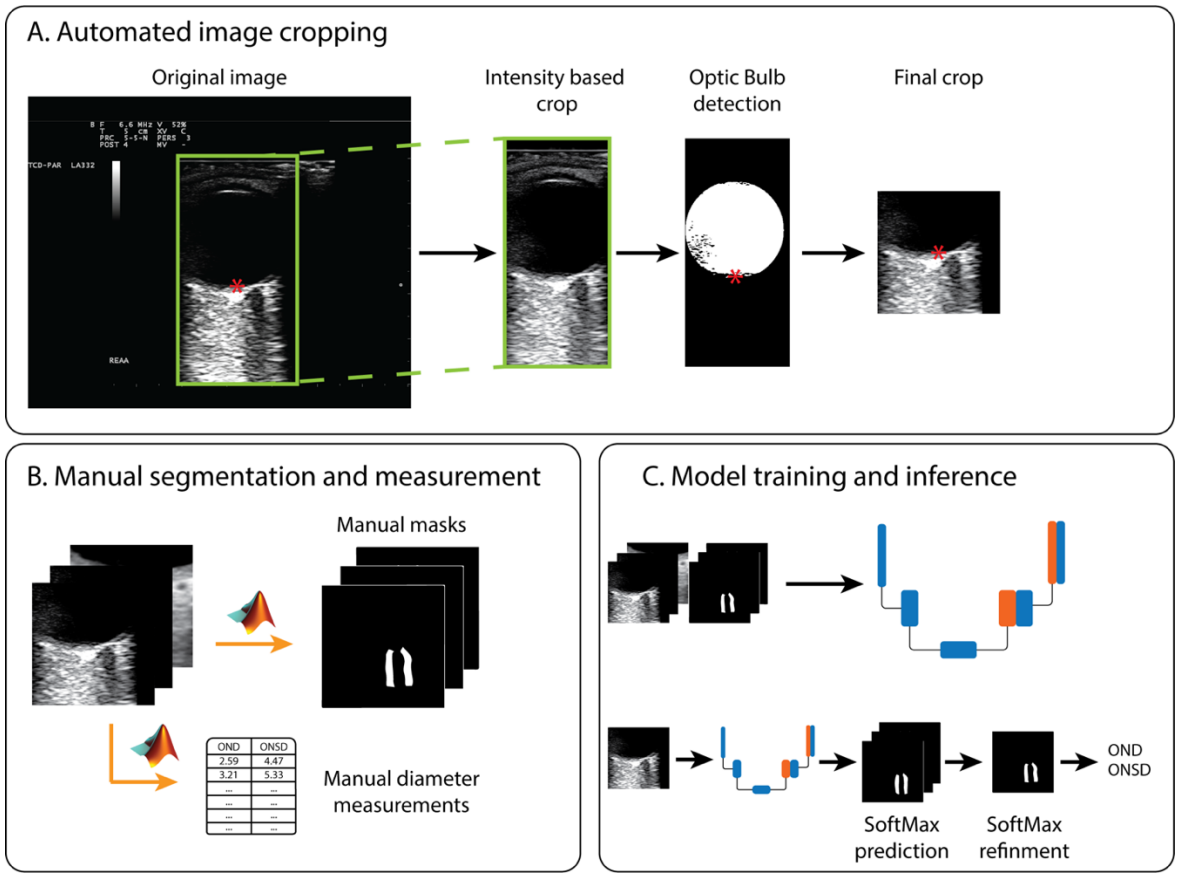


Figure 10 In panel A is displayed the image cropping process. Panel B illustrates the usage of the two GUIs to extract masks and diameters. Panel C shows UNet training and inference steps.

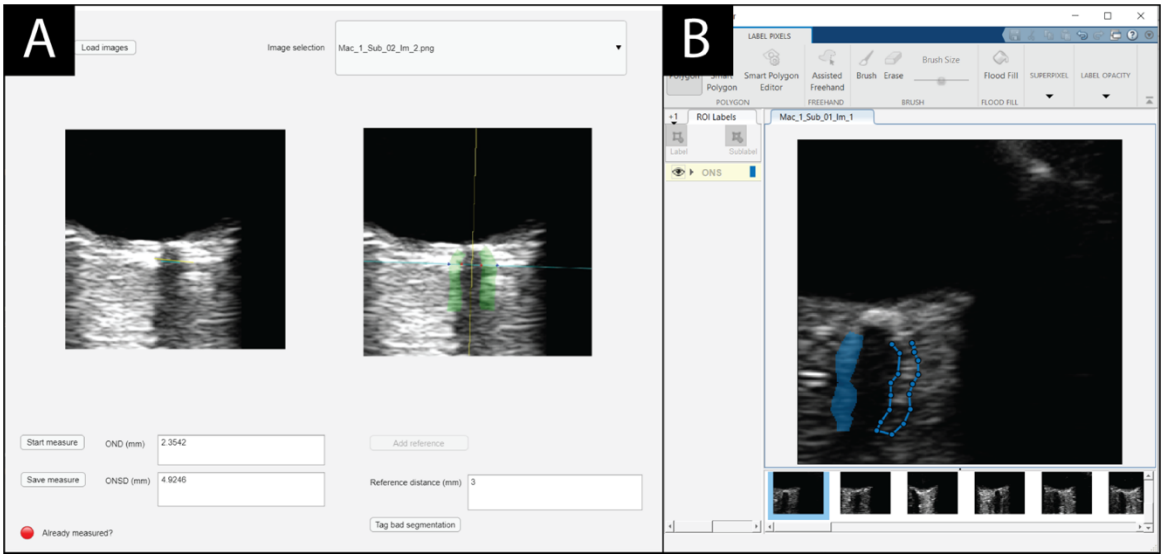


Figure 11 A) Custom User Interface for diameter measurement. On the left panel, the operator can measure OND and ONSD by tracking two lines. On the right panel, there is a guide that highlights the 3 mm zone. B) ImageLabeler User Interface for ONS segmentation.

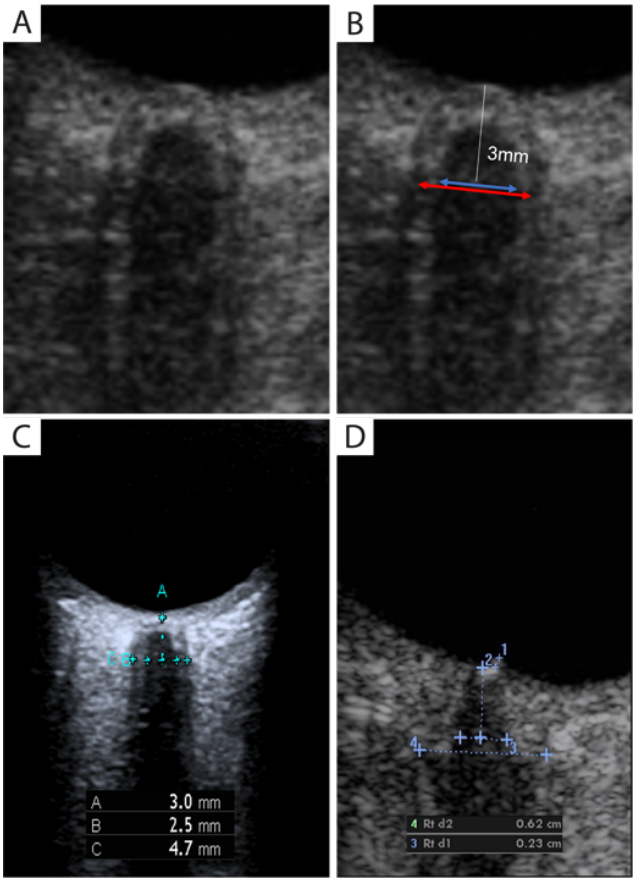


Figure 12 A) Depiction of the optic nerve and its meningeal sheath. B) OND (blue arrow) and ONSD (red arrow) measured 3mm behind the papilla. C) Measurements of the OND and ONSD in a healthy subject. D) Measurements for a pathological subject.

4.4. Model Training and Inference

A segmentation model was trained using a UNet with 5 down-sampling stages and an encoder based on a Resnet50 pretrained on ImageNet. The objective function employed the Dice loss, optimized through Adam with an initial learning rate of 0.0002. The training was done with a batch size of 32 images. To identify the optimal model, Dice Loss was monitored on a Validation set throughout the training phase. The learning rate was reduced by half at the 20th and 40th epochs. If there wasn't a minimum 1% enhancement in the Validation loss compared to the best epoch after 10 consecutive epochs, training was halted. Input images and masks underwent data augmentation to stabilize the training process. Each batch extracted from the dataset had a set probability (p) for undergoing specific transformations, which were executed in a randomized order: Horizontal flip (p = 0.5), Scaling and Rotation (scale range of ± 10%, rotation range of ±10°, p = 1), Gaussian Noise (p = 0.3), CLAHE (p = 0.3), Noise multiplication (p = 0.2), Optical alterations (p = 0.2), Image sharpening (p = 0.25), and Image blurring (p = 0.5). The code offers further insights. Pytorch along with the Albumentations and Segmentation Models libraries were used to develop this method [134], [136], [171].

The limited data size posed challenges for implementing a deep learning model. Therefore, a 5-fold Cross-Validation (CV) was integrated to ensure reliable performance evaluation. To compare with the ground truth, the 5 test folds were concatenated, representing the entire dataset. In each iteration, 3 subsets functioned as training data, and one each for validation and testing. To enhance the ON and ONSD segmentation, this 5-fold CV procedure was repeated 8 times to produce an Ensemble prediction. This was feasible thanks to the rapid training time, with each session finalized in under 20 minutes with an RTX 3060, leading to a 12-hour total training duration.

For the Ensemble prediction, every SoftMax output was binary-coded using a 0.5 threshold, generating segmentation masks. Each mask was then subjected to morphological processes to rectify gaps, and smooth edges (specifically, a disk structural element with a 3-pixel radius for opening, followed by a 5-pixel radius disk for closing), and to maintain only the two largest connected objects in the mask. The eight refined masks were averaged to produce a probability map, which was then binarized using a 0.6 threshold. Identical morphological interventions used on individual masks were employed to refine the projected region. Both threshold values and structural element dimensions were determined based on empirical observations. The comprehensive post-processing steps are elaborated in Polito’s BioLab open repository. A visual representation of the entire data workflow can be viewed in Figure 10, while Figure 13 offers an in-depth visual explanation of the morphological operations on the SoftMax outputs and their average.

Additionally, to further validate our model's generalization capability, the UNet model was trained using images from three devices, with images from the fourth device serving as an external testing set. This approach also utilized a 5-fold CV, without repetitions. Four folds were designated for training, and one for validation. The test set comprised images from the excluded device. As a result, each test set offered five distinct predictions, which were averaged in the same manner as described earlier.

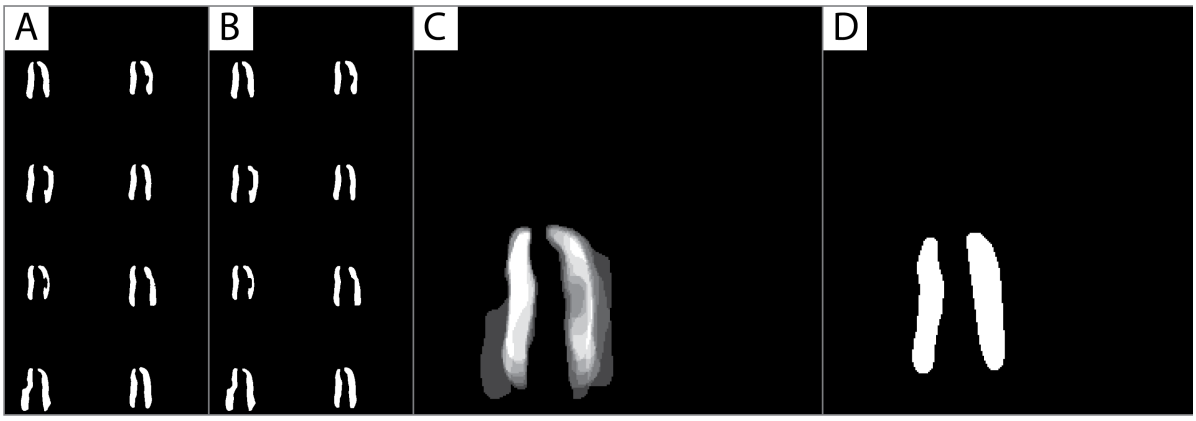


Figure 13 A) Binarized SoftMax outputs from 8 separately trained models. B) Post-processed SoftMax. C) Averaging of the 8 SoftMax results in a probability map. D) Post-processing of the probability map (thresholding at 0.6 and morphological operations).

4.5. Model evaluation

To assess model's effectiveness, the Dice score and the Hausdorff Distance were computed by comparing the manually created masks to the automatically generated ONS masks. The evaluations were confined to the image regions where manual masks were present. This restriction was necessary since the automated system identified the ONS even in low SNR areas, whereas manual segmentation was exclusive to high SNR regions. Additionally, a paired Wilcoxon test was performed to determine if the metric distributions from individual runs significantly deviated from those of the Ensemble prediction.

The primary objective of this methodology is to accurately gauge the OND and ONSD. Diameters obtained automatically were compared with the benchmark (i.e., manual segmentations by A1). Three distinct methodologies were employed for these comparisons: the first method involved diameter assessments by another analyst (A2) employing the same GUI as the benchmark; the second derived diameters from manual masks (referred to as A2mask); and the third estimated diameters from masks generated by the network (UNet). These three techniques were evaluated by calculating the absolute error (as given in Eq. 1), ICC (Intra Class Correlation) (2,1), Bland-Altman plots, and correlation plots.

$$AbsoluteError_{Diameter} = |Diameter_{Method} - Diameter_{AN}| \quad (1)$$

This research seeks to offer recommendations for capturing TOS images that enable automated systems to achieve human-comparable results. As part of this, the segmentation and diameter measurement outcomes were compared across varying SNR levels. SNR was defined as $SNR = 20 \times log_{10}\frac{P_s}{P_n}$ with P_s as signal power and P_n as noise power. The SNR was calculated in two unique locations using the labels manually set by A2. Initially, signal power was estimated within the ONS labels, and noise power within the ON label. Subsequently, while the signal power remained within the ONS labels, the noise power was derived from the retrobulbar fat region encircling the ONS labels. Figure 14 shows these two measurements. Three SNR categories were delineated:

- Low SNR: $SNR_i < \overline{SNR} - \sigma_{SNR}$
- Average SNR: $\overline{SNR} - \sigma_{SNR} > SNR_i < \overline{SNR} + \sigma_{SNR}$
- High SNR: $SNR_i > \overline{SNR} + \sigma_{SNR}$

The effect of SNR on segmentation was assessed by comparing the Dice score across different SNR tiers. To discern the influence of SNR on diameter evaluations, the SNR value distribution was studied based on the accuracy of the OND and ONSD calculations. Measurements within a 0.2 mm range from the Ground Truth were designated as "correct", while those beyond this threshold were labeled as either "over" or "under" estimates. The SNR value distribution for each segment was subsequently examined.

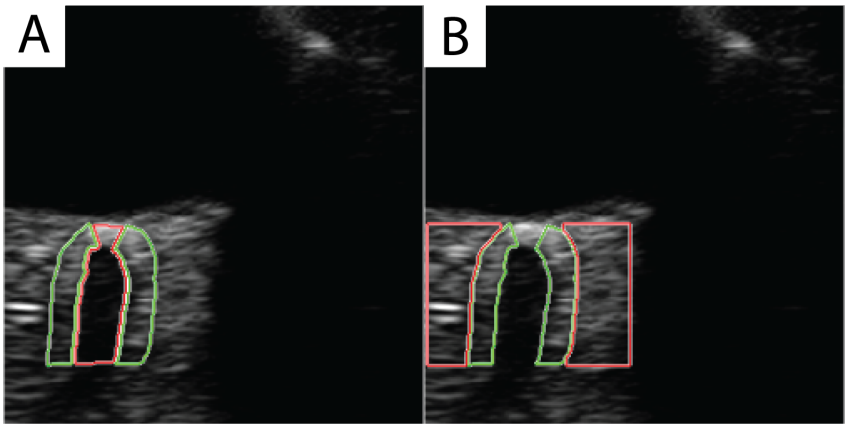


Figure 14 Areas considered during Signal and Noise Power calculation. A) Signal power (green label) is calculated in the ONS area, while Noise power (red label) is in the ON area. B) Signal Power is calculated in the ONS area and Noise power in the retrobulbar fat.

4.6. Results

For each repetition, the Dice score average was between 0.695 and 0.713 with a standard deviation ranging from 0.141 to 0.147. In comparison, the Ensemble prediction's average Dice score was 0.719 with a standard deviation of 0.139. The Wilcoxon tests consistently indicated a significant increase ($p < 0.01$) between the Ensemble prediction and the individual run Dice scores. In parallel, the average Hausdorff Distance (HD) for each iteration varied between 8.9 and 9.3 pixels, having a standard deviation between 4.0 to 4.4 pixels. The Ensemble's prediction, however, noted a lower HD of 8.2 ± 0.5 pixels, significantly differing from the individual run outcomes ($p < 0.01$). Figure 15 visualizes the Dice score and HD for each iteration and the Ensemble prediction through boxplots.

In terms of the network's adaptability to different devices used as test sets, the Dice scores were: 0.658 (± 0.156) for Machine1 (MyLab Esaote Homburg, N = 207), 0.726 (± 0.121) for Machine2 (MyLab Esaote Turin, N = 113), 0.612 (± 0.157) for Machine3 (Toshiba Aplio 300, N = 93), and 0.689 (± 0.112) for Machine4 (GE-Vivid7, N = 51). Figure 16 showcases the Dice Score and HD for these tests using boxplots.

During manual diameter assessments, A1 deemed 45 images unsuitable for OND and ONSD measurements due to inadequate SNR. The image exclusions per device are detailed in Table 8. However, these images remained in the segmentation performance evaluation since A2 annotated all images. Table 8 shows the results of diameter comparisons to the benchmark A1.

The automatic method slightly underestimates OND, with an average error of -0.253 mm (MAE = 0.323±0.246 mm, MSE = 0.165±0.273 mm) and a Pearson's correlation value of 0.69. Correlation values for A2 to A1 and A1 to A2 mask were 0.85 and 0.84, respectively. The ONSD measure presented a negligible mean bias of 0.027 mm (MAE = 0.394±0.320 mm, MSE = 0.257±0.426 mm). Correlation values for A1 with UNet, A2mask, and A2 were 0.69, 0.86, and 0.89 respectively. Bland-Altman plots didn't exhibit any discernible patterns in OND error distribution. However, ONSD automatic

Pool	name	OND Error (mm)	ONSD Error (mm)	OND Abs.Err.(mm)	ONSD Abs.Err.(mm)	iccOND	iccONSD	Low Quality
Full	$A2_{mask}$	0.193±0.244	0.068±0.368	0.244±0.194	0.288±0.238	0.752	0.861	45
	UNet	0.253±0.318	0.027±0.507	0.323±0.246	0.394±0.319	0.518	0.661	
	A2	0.051±0.226	0.166±0.337	0.180±0.146	0.293±0.234	0.845	0.860	
Homburg MyLab Esaote	$A2_{mask}$	0.211±0.249	0.091±0.411	0.260±0.197	0.321±0.272	0.631	0.810	20
	UNet	0.261±0.291	0.026±0.510	0.324±0.218	0.410±0.303	0.402	0.610	
	A2	0.046±0.226	0.226±0.297	0.178±0.147	0.295±0.229	0.770	0.844	
Turin MyLab Esaote	$A2_{mask}$	0.158±0.196	0.117±0.315	0.196±0.157	0.268±0.201	0.840	0.873	8
	UNet	0.229±0.267	0.083±0.440	0.281±0.212	0.332±0.298	0.618	0.710	
	A2	0.030±0.219	0.161±0.411	0.167±0.144	0.345±0.273	0.877	0.794	
Toshiba Aplio 300	$A2_{mask}$	0.196±0.281	0.023±0.359	0.249±0.234	0.281±0.222	0.752	0.885	15
	UNet	0.257±0.467	0.022±0.644	0.413±0.335	0.479±0.428	0.334	0.559	
	A2	0.079±0.275	0.117±0.363	0.225±0.176	0.303±0.229	0.818	0.876	
GE Vivid7	$A2_{mask}$	0.195±0.256	0.033±0.290	0.275±0.166	0.220±0.189	0.653	0.913	2
	UNet	0.260±0.229	0.026±0.436	0.274±0.211	0.350±0.257	0.503	0.714	
	A2	0.087±0.170	0.047±0.218	0.156±0.110	0.178±0.132	0.856	0.947	

Table 8 Evaluation of the measurement errors with respect to the ground truth of the compared methods. A2 is the manual diameter measurement from the second operator. UNet refers to the measurements made on the predicted masks. A2mask refers to the measurements made on the predicted masks. A2mask refers to the measurements made on the manual masks. Other abbreviations: OND (Optic Nerve Diameter), ONSD (Optic Nerve Sheaths Diameter), icc (Intraclass Correlation Coefficient).

measurements tended to overestimate as manual ONSD increased. Correlation plots didn't highlight any concerns regarding error deviation uniformity for OND and ONSD. Figure 17 depicts these plots for both OND and ONSD.

Figure 18.A contrasts the Dice scores for the left and right ON sheaths and the entire image using boxplots, segmented by SNR category, displaying the Dice score on the y-axis. These visuals show the impact of SNR levels on Dice distribution, indicating potential outliers and errors even in high SNR images.

Figure 18.B compares SNR values with OND measurement accuracy. Grouped by measurement accuracy categories, the y-axis displays the SNR levels. A marginally elevated SNR distribution is evident for accurate OND values. Analogous observations emerge from Figure 18.C, where ONSD measurements are compared to SNR levels.

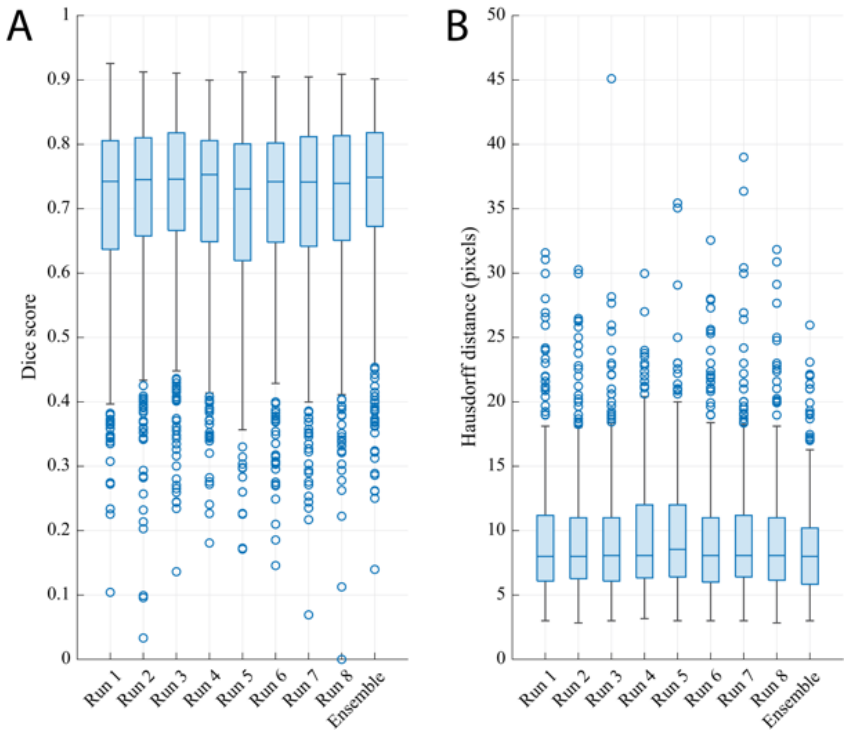


Figure 15 A) Boxplots showing the Dice scores for the 8 runs and the ones resulting from the Ensemble of the SoftMax. B) Boxplots of Hausdorff distances in pixels between the ONS contours of the ground truth and the automated masks for the 8 runs and the Ensemble.

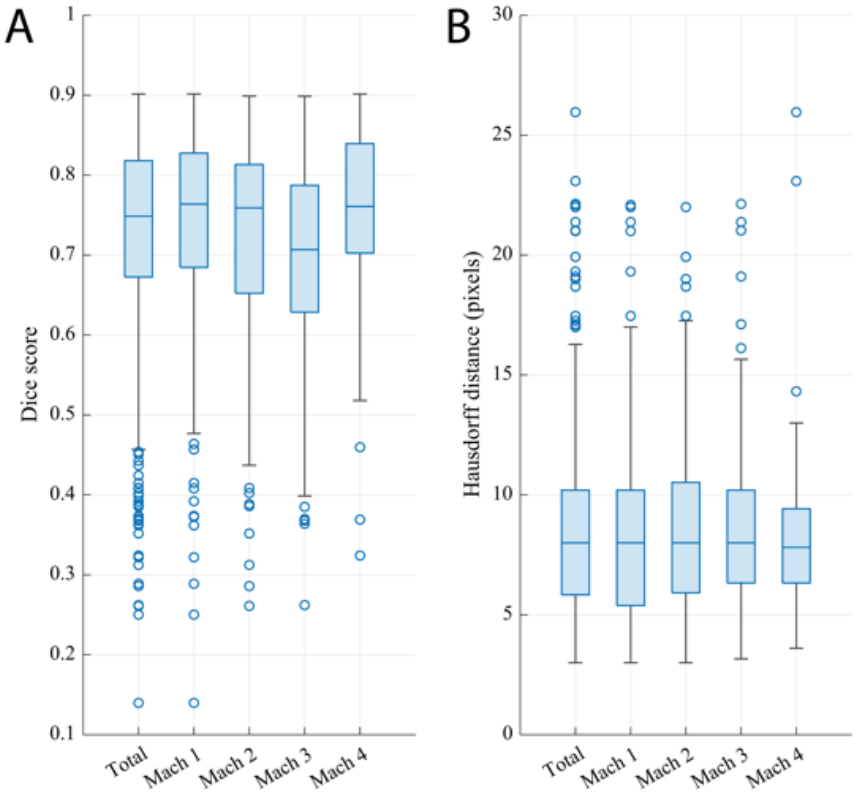


Figure 16 Dice scores for the subsets when that machine was used as the test set. First column is the union of the four runs. B) Hausdorff distances. Mach 1 (Esaote), Mach 2 (Esaote), Mach 3 (Aplio300 and AplioXG), Mach 4 (Vivid7).

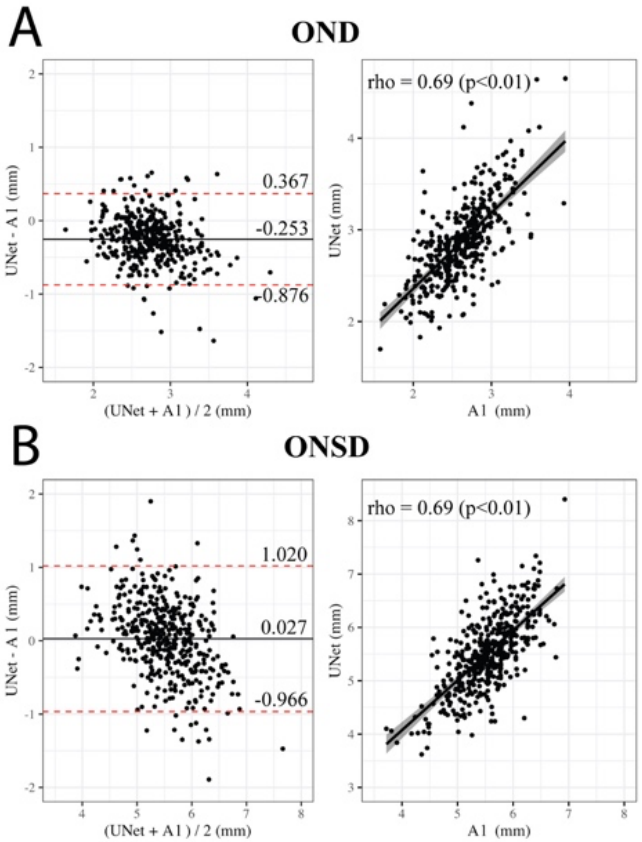


Figure 17 A) Bland-Altman and Correlation plots. B) ONSD. The red dashed lines are at 1.96 SD. rho is calculated using Pearson's correlation. A1: manual measurements obtained by the most experienced operator.

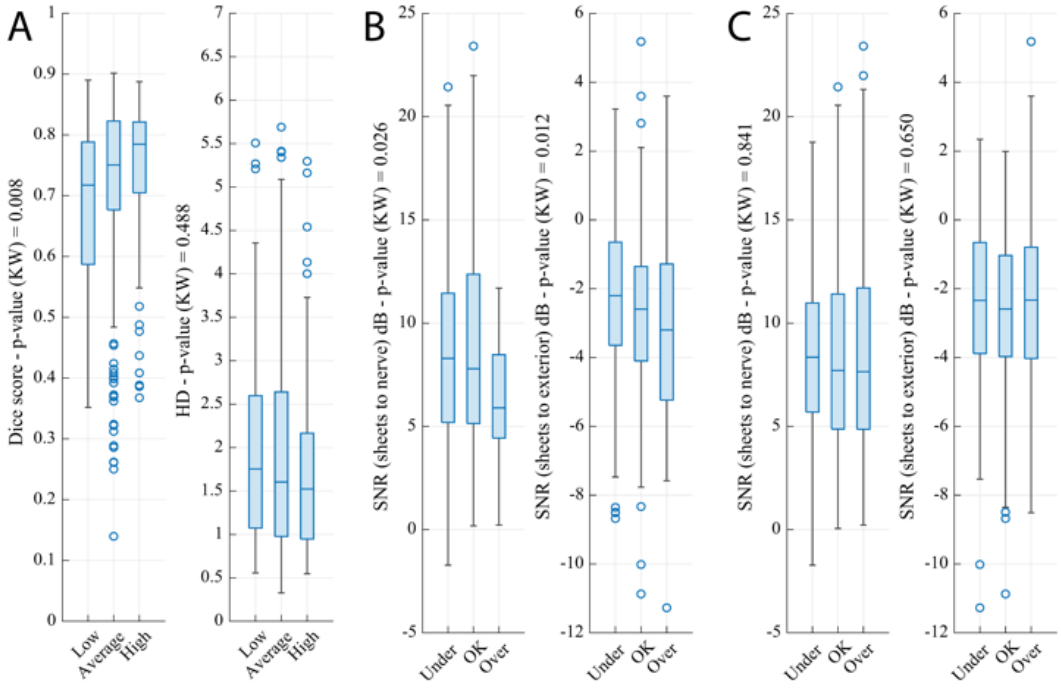


Figure 18 A) Boxplots visualizing Dice Score and Hausdorff Distance distributions for distinct categories of SNR. B) These Boxplots visualize SNR distribution depending on the correctness of the OND measurement (GT ± 0.2 mm). SNR is calculated first from the ON (left), then from the ON sheaths to their peripherical area(right). C) SNR distribution depending on the correctness of ONSD measurement.

4.7. Discussion

In this study, a deep learning (DL)-driven segmentation technique was introduced offering promising outcomes in detecting the optic nerve and its surrounding sheaths in TOS imagery. Technically, this method leverages a unique refinement operation, which averages predictions from multiple models and employs basic morphological processes. This enhancement aids the system in dealing with the inherent low SNR of ultrasound images. Figure 13 illustrates this; single network outputs (panels A and B) might depict unnatural shapes and segmentation inconsistencies. However, the ensemble prediction, represented in panels C and D, rectifies these irregularities, resulting in enhanced segmentation fidelity. While this approach, combined with a 5-fold CV, extends the training duration, it brings improvements to both the segmentation and validation phases. Ideally, a substantially larger dataset could potentially bypass both CV and multi-output averaging, but given our study's dataset size, these steps were deemed essential. This approach could be expanded by monitoring also the uncertainty of the prediction. Methods such as Montecarlo sampling or a simpler Dropout selection could potentially replace the multiple training of the network. Moreover, providing the physician with an uncertainty heatmap could help in bringing this method closer to clinical applicability.

Comparing diameters from manual masks to our benchmark, the OND and ONSD measurements reaffirmed consistent inter-observer concordance as documented in prior studies. To match the inter-rater variability range, the DL system's precision needs to be improved. Systems where DL-driven segmentation tools applied to ultrasound images attain Dice scores surpassing 0.8 and exhibit an agreement with the benchmark akin to intra-operator variability have been previously documented [28, 29, 30, 31]. Our relatively lower performance can be attributed to our dataset's limited size and complexity, encompassing multi-center data, diverse devices, and a broad ONDS diameter range (3.62 mm to 8.40 mm). Nevertheless, the outcomes align with expectations and can potentially be scaled with extended training data or reduced dataset diversity.

The network's generalization ability was scrutinized via cross-validation folds and by excluding one machine during training. The former consistently yielded reliable results, suggesting that the network can adeptly segment images like its training dataset. However, the latter, emphasizing device-to-device generalization, was more demanding, particularly for Machine 3. This underscores the method's capacity to adapt across different devices but highlights the need for specific tuning per device for optimal alignment with expert annotations. This requirement stems from varied settings and image enhancement algorithms across device manufacturers, impacting the ultrasound image's overall look and texture.

Correlation and Bland-Altman plots revealed a correlation between the automated and benchmark OND and ONSD measurements. However, these coefficients are slightly inferior to those derived from another manual operator and those from manual masks. This observation mirrors the segmentation performance, emphasizing this methodology's constraints. The ICC values in Table 8 further validate this limitation, indicating that automated measurements slightly deviate from the benchmark compared to another manual operator.

Achieving consistency in diameter measurement is challenging due to the mm/pixel conversion factor and ultrasound images' lateral resolution. In our collection, the median conversion factor stands at 0.1061 mm/pixel, with an ONSD absolute error standard deviation for Unet measurements being 0.507 mm. Consequently, minor pixel errors can compromise measurement precision. The ultrasound images' lateral resolution oscillates between 0.47 mm to 0.60 mm, situating the ONSD absolute error standard deviation within this bracket. This method's uncertainty margin closely mirrors the inherent physical constraints of ultrasound. To elevate accuracy, enhancements in ultrasound signal and image

processing are imperative. For the signal processing and image reconstruction phase, innovative DL-driven techniques are under exploration [11], alongside super-resolution ultrasound methods [172]. From the image processing angle, future endeavors might involve heightening image resolution via super-resolution algorithms based on GANs or Diffusion Models.

Both the Dice score and OND/ONSD correlation analysis yielded analogous conclusions. It's important to understand that segmentation approaches and automated systems for ON and ONS analysis necessitate distinct evaluations. While the Dice score is useful for comparing various segmentation techniques on an identical dataset, correlation analysis is optimal for gauging a system's clinical efficiency. The SNR analysis manifests a correlation between SNR level and segmentation performance, but not diameter measurement accuracy. This discrepancy can be reasoned by the inherent challenges of segmentation, especially in more demanding regions. Elevated SNR aids in sheath detection, particularly in these challenging zones. However, the 3 mm area typically has a higher SNR, minimizing its influence on automated diameter accuracy. For optimal automated algorithm performance, acquisition operators should prioritize clear contrast between the ON and ONS and well-defined ONS boundaries relative to surrounding structures. An accurate, full-length ONS segmentation system might pave the way for exploring ON and ONS diameter variability based on measurement depth.

In conclusion, enlarging or reducing dataset variability should improve both segmentation and diameter measurement precision. Augmenting the sample count would refine this method's generalization, facilitating the creation of a foundational model transferable across multiple devices and protocols. Subsequently, tailoring this model to a standardized, less diverse dataset should further enhance accuracy. Merging these advancements would lead to a system proficient in ONS segmentation and OND and ONSD measurement.

On the broader picture of this thesis, this chapter showed an application where DL-based system still fails to reach human performance highlighting the importance of developing better representations of medical images and ultrasound images in particular. In this application, the complexity of the model that could be applied was limited by the size of the dataset. To improve the performances, a system based on an ensemble of models was implemented, still the correlation between the automatic measurement and the benchmark was lower than inter-rater variability. When dealing with datasets so challenging one option could be fine-tuning foundational models, but application of such technique in ultrasound is still underperforming compared to training dedicated models to specific applications. Another approach could be standardization of image data and augmentation of dataset using generative models. The approach proposed in Chapter 6 will go in this direction.

In the next chapter the opposite situation will be explored where, having a very large dataset, the capacity of the network was scaled up to learn a diverse and complex task and to improve the segmentation performance.

5. Deep learning and radiomics for quantitative muscle ultrasound analysis

5.1. Introduction

Muscle ultrasound (MUS) using B-mode imaging has become an important tool in the study of the musculoskeletal system, offering significant insights into muscle architecture, morphology, and composition [173].

One of the key strengths of muscle ultrasound lies in its ability to evaluate various muscle tissue properties beyond just echogenicity. It can assess muscle atrophy, vascularization, and elasticity, proving particularly useful in the diagnosis of neuromuscular diseases evaluating fibrosis and fatty replacement. Moreover, muscle ultrasound can capture dynamic muscle movements, such as fasciculations and fibrillations, which enhances its sensitivity for detecting conditions like ALS (Amyotrophic Lateral Sclerosis) [174] and age-induced sarcopenia [175].

The qualitative or semi-quantitative evaluation of US images plays a foundational role. Traditional assessments often involve visual or semi-quantitative analysis, such as the Heckmatt scale [176], which rates muscle echogenicity to estimate muscle condition. While allowing a comprehensive view of muscle health they are limited by their dependence on user interpretation and the variability in ultrasound machine settings [174].

Quantitative muscle ultrasound (QMUS) techniques offer a more objective and detailed assessment of muscle tissue, extending the capabilities of traditional ultrasound imaging. The implementation and effectiveness of QMUS in clinical practice face the challenge of device dependency, requiring specific calibration and reference values for each ultrasound machine used [177]. This lack of standardization complicates the comparison of results across different machines and settings. Furthermore, patient-specific characteristics such as age and obesity can also influence the ultrasound results, adding complexity to the interpretation of findings [178].

Muscle echogenicity (EI) and other first-order statistical features of ultrasound imaging are commonly used as a proxy for muscle composition, indicating the degree of adipose and connective tissue infiltration within muscle tissue. However, the limitation of first-order features lies in their focus on individual pixel intensity values, overlooking the spatial relationships between pixels [179]. Texture analysis techniques based on higher order features such as Haralick [180] and Galloway features [181] provide a more comprehensive evaluation of muscle tissue revealing gray-level patterns and interrelationships between pixels.

Studies have demonstrated the potential of GLCM in offering better discriminatory value for characterizing muscle composition compared to first-order features. This approach is less influenced by participant characteristics and ultrasound equipment settings, making it more versatile and reliable for clinical and research applications. GLCM's ability to distinguish between muscles with different architectural characteristics has been demonstrated by Molinari et al. [182]. It was found particularly valuable in studies focusing on quadriceps muscles [183], in monitoring disease diagnosis [184] and progression [185]. Only certain GLCM variables showed correlations with muscle strength, suggesting that while GLCM can discriminate muscle composition, its relationship with muscle force is not always straightforward [183], [186]. Aside from Haralick features, local binary pattern (LBP) operator has been used for the assessment of spastic muscles [187].

To perform quantitative analysis, the first step is to define the Region of Interest (ROI) where to measure the features involving a segmentation step. Automating this segmentation step could potentially reduce the workload on physicians and reduce variability related to the different readers for large scale studies. The unique challenges of muscle segmentation, including heterogeneous textures, variable fiber orientations, and dynamic changes, necessitate adaptable algorithms specific to each application.

While previous works focused primarily on segmentation or feature analysis, in this chapter the aim is to give a holistic view of the MUS automated analysis including all the steps, from the initial classification and segmentation to the assessment of muscle condition by a modified Heckmatt scoring using texture analysis. Moreover, in current literature, automatic methods are limited to analyzing only few muscle groups, with facial muscles totally excluded in this analysis. The method presented in this chapter performs automatic segmentation and classification for 16 different muscles, paving the way for automatic and reproducible assessment of subject condition for clinical adoption. Therefore, in this chapter, an end-to-end automatic pipeline for muscle classification, segmentation, and quantitative assessment is presented. This proposal stems from our first approaches with deep learning for muscle segmentation in deep learning [188], [189]. The research presented in this chapter expands the application of DL methods adding automatic classification of muscle type and implementing an analysis of texture features that mimics the actual evaluation performed by a physician.

The state of the art of automatic segmentation of MUS will be presented, then the dataset used to develop the pipeline will be described, as well as the methods and the results. The feature extraction step will be described alongside the ML model for Heckmatt scoring. The chapter will finish with a discussion evaluating the potentialities and limitations of this approach.

5.2. State of the art – automatic segmentation

The evolution of automatic MUS segmentation has transitioned from rule-based methods to deep learning-based segmentation techniques. This shift was urged by the unique challenges posed by muscle segmentation in comparison to other organs like the breast or cardiovascular structures in ultrasound imaging. Muscle tissues are inherently complex, featuring heterogeneous textures, variable fiber orientations, and often blurry boundaries, unlike more uniform organs such as the liver or cardiovascular structures. This anatomical complexity makes the development of precise segmentation algorithms particularly challenging. Additionally, the dynamic nature of muscles, which alter shape and appearance with movement, further complicates the segmentation process. Automated techniques must, therefore, be exceptionally adaptable to effectively manage these variations. Another hurdle in muscle ultrasound imaging is the typically low contrast and resolution, which stems from the similar echogenicity of muscle tissue and adjacent structures, complicating the distinction of muscle boundaries. Moreover, the scarcity of large, annotated datasets for muscle ultrasound impedes the progress in developing and refining deep learning-based segmentation techniques, a challenge less pronounced in the imaging of other organs.

One of the first methods for real-time segmentation of cervical muscles, crucial for conditions like Cervical Dystonia, was developed by integrating ultrasound with MRI data through multimodal registration based on shape statistics [190]. In the research presented in [191], the focus was set on leveraging the cohesiveness of fascicle orientation for image segmentation, where fascicle orientations were tracked using the Lucas-Kanade optical flow algorithm on manually predefined points. The application of Radon transform and wavelet analysis, as explored in [192], was instrumental in assessing variations in fascicle curvature.

The development of the MUSA algorithm [193] introduced fully automated segmentation for longitudinal ultrasound images of four different muscle types, employing multiscale filtering for aponeurosis selection. Building on this, the TRAMA algorithm [194] extended these capabilities to transverse images, enabling automated measurement of muscle cross-sectional area (CSA).

Deep learning-based models have revolutionized the automatic assessment of more challenging structures, such as muscle junctions or pathological muscles with elevated echogenicity, which often have less distinguishable boundaries compared to healthy tissues. A notable advancement was the application of region-based convolutional neural networks (RCNN) for localizing myotendinous junctions (MTJ) in consecutive ultrasound images, as seen in [195]. The initial use of the UNet model, reported in [189], marked a significant step in considering echogenicity differences in muscle for algorithm evaluation. Subsequent improvements in the UNet architecture, such as those discussed in [196] and [197], introduced modules for attention and feature fusion to enhance segmentation accuracy. SCCNet [142] specifically aimed to tackle the challenges of segmenting high-variability ultrasound images, like those of neck muscles, using a self-correction boundary module and a class-context module. However, the recent exploration of fine-tuning foundational models like the Segment Anything Model (SAM)[198] for vision applications by Mazurowski et al. [199] revealed limited success in transverse ultrasound images, reaffirming the complexity of muscle segmentation in ultrasound.

5.3. Dataset description

To the best of our knowledge, the dataset used to develop the current approach is the largest dataset published so far and will be made open-access as soon as the related paper is published.

This dataset comprises multi-site acquisition from 290 subjects from 21 different muscles. For each muscle multiple images are available. Muscle ultrasound was performed with an Esaote MyLabTwice ultrasound scanner (Esaote SpA, Genoa, Italy) using a LA533 3–13 MHz linear transducer with an axial resolution of around 0.2 mm. System presets remained unchanged across all muscle measurements to maintain consistency. Generally, a 4 cm depth preset was applied, positioning the focal area at the image's lower edge to prevent uneven grayscale distribution in this ultrasound system. Exceptions were made for thigh and abdominal muscles, where a 6 cm depth setting was used, again with the focus at the lower edge. A uniform gain setting of 50% was employed for all muscles. Scanning was conducted in the transverse plane at a specific anatomical site representing the maximal muscle bulk, without permitting image modifications during or post-scanning. Patients underwent examination in a supine position with limbs in a restful anatomical stance or seated in a wheelchair for non-ambulatory individuals. Muscle ultrasound assessments involved visual analysis using the 4-point Heckmatt grading scale, conducted blind to clinical, EMG, and previous ultrasound/EMG data and laboratory results. Quantitative analysis compared measured gray levels against predicted values based on age, sex, and BMI, yielding a z-score. Z-scores ≥ 2, equivalent to ≥ 95th percentile, and Heckmatt grades ≥ 2 were categorized as abnormal. Images were collected during clinical practice and deidentified before sharing between centers so no further ethical approval was deemed necessary. Muscle annotation was performed during clinical routine by expert neurodiagnostic technicians in a strictly protocolized to guarantee inter-observer consistency and used as the target of the segmentation network. 25005 B-mode images were acquired from 290 subjects in 21 different muscles. For 110 subjects the Heckmatt semi-quantitative grading was available for 7 different muscles. Table 9 presents a summary of subjects' characteristics, while Table 10 presents a description of the distribution of the different muscle types present in the dataset. For each image only one muscle has been annotated with a segmentation map and a label describing the type of muscle in the image have been assigned. For the development of the segmentation and classification network we

chose to retain only muscles with at least 200 images to avoid having a too unbalanced dataset. Consequently, images of Extensor digitorum brevis, Flexor carpi radialis, Flexor digitorum profundus, Levator labii superior, and Peroneus tertius were removed from the dataset. The resulting class distribution is available in Figure 19.

Characteristic	Value
Total count	110
Sex (0: Male, 1: Female)	58 (0) 52 (1)
Age (years)	52.71 +/- 14.25
Weight (kg)	79.77 +/- 13.57
Height (m)	1.76 +/- 0.09
BMI	25.82 +/- 3.91

Table 9 Subjects' characteristics with semi-quantitative evaluation.

Muscle	Number	MGL mean	MGL std	Area mean	Area std
Biceps brachii	1740	82.813	23.505	18.392	5.145
Deltoideus	1030	73.096	15.019	17.686	4.095
Depressor anguli oris	1500	56.054	15.275	0.625	0.268
Digastricus	1449	46.631	10.054	3.019	1.047
Extensor digitorum brevis	58	76.390	16.840	0.833	0.307
Flexor carpi radialis	6	54.843	1.568	6.409	0.312
Flexor digitorum profundus	6	67.271	3.845	12.673	1.618
Gastrocnemius medial head	1746	81.446	18.657	16.937	4.113
Geniohyoideus	732	37.522	11.995	5.009	1.657
Levator labii superior	17	92.159	12.099	0.683	0.342
Masseter	1479	67.214	14.557	10.882	2.575
Mentalis	1244	51.475	12.625	0.843	0.554
Orbicularis oris	1341	65.367	14.137	0.463	0.190
Peroneus tertius	162	79.487	19.771	10.461	3.362
Rectus abdominis	1737	97.887	20.615	8.091	2.706
Rectus femoris	1737	90.269	23.001	15.197	4.923
Temporalis	1472	79.576	10.772	14.864	4.482
Tibialis anterior	1736	86.661	17.132	16.695	5.319
Trapezius	1521	100.309	22.257	7.545	3.474
Vastus lateralis	1723	78.804	17.411	18.707	5.324
Zygomaticus	2569	96.580	17.522	0.840	0.561

Table 10 Dataset muscle type distribution. MGL: mean gray level inside CSA, range 0-255. Area of CSA expressed as cm^2

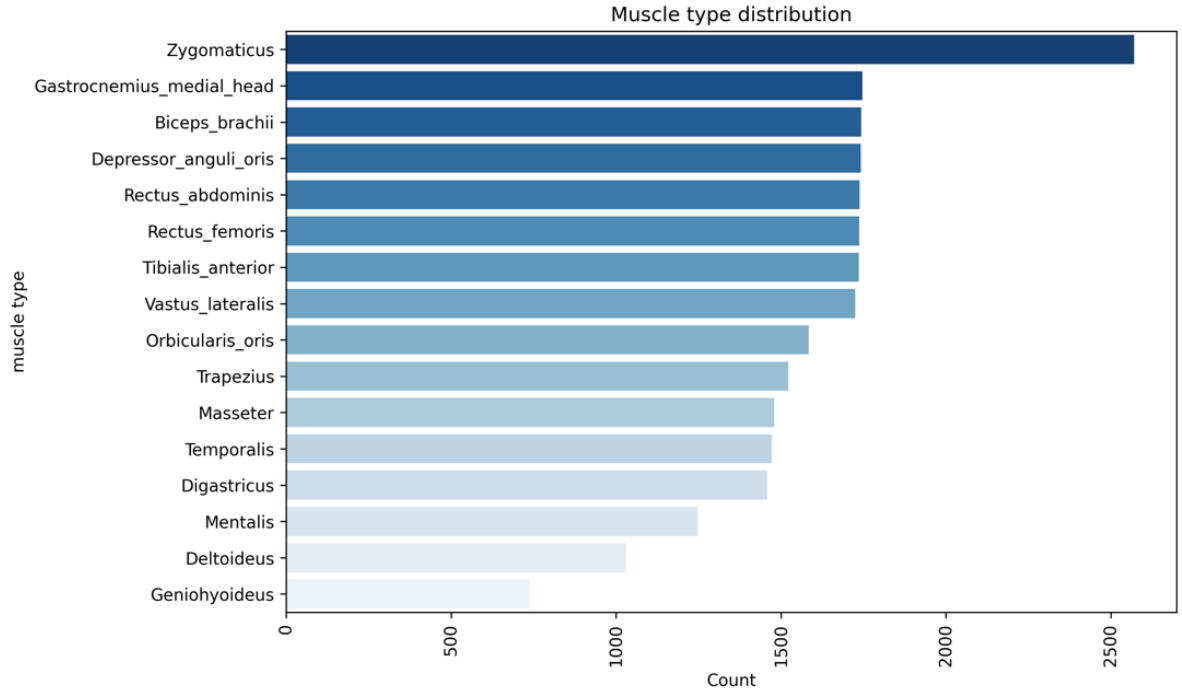


Figure 19 Dataset muscle type distribution after selection. Zygomaticus presents a higher count since we grouped Zygomaticus major and Zygomaticus minor images under physician evaluation give their similar appearance and physiology.

5.4. Segmentation model description

The segmentation process faces unique challenges: muscles exhibit varying textures and fiber orientations, and their boundaries with adjacent tissues can be indistinct. To effectively tackle these challenges, Knet architecture [200] was used in integration with the UPerNet (Unified Perceptual Parsing for Scene Understanding) [144] framework and employing Swin transformer [201] as its feature extractor. A schematic description of the model is available in Figure 20. The network was implemented using OpenMMlab implementation and development framework [202].

With respect to a more standard approach like UNet, the K-Net architecture improves the segmentation performance by updating conditional kernels based on the class of each detected object. This results in a more precise mask prediction, tailored to the unique characteristics of specific muscle tissues in ultrasound images. These conditional kernels are a type of convolutional kernel that dynamically adjusts based on the input data. Unlike standard kernels that apply uniform weights, conditional kernels in Knet adapt their weights according to the specific features of the muscle tissue being analyzed.

The model employs a Swin transformer as its feature extractor, providing a robust and precise representation of the complex muscle structures. The Swin transformer is particularly effective in constructing hierarchical feature maps for muscle ultrasound using information at multiple scales. After feature extraction, the UPerNet leverages global and local context information by integrating features across various pyramid scales. This approach combines a bottom-up pathway, capturing the low-level features of muscle textures, with a top-down pathway that refines the segmentation map using skip connections. The lateral connection between these pathways merges feature maps of differing resolutions, thereby enhancing the segmentation accuracy.

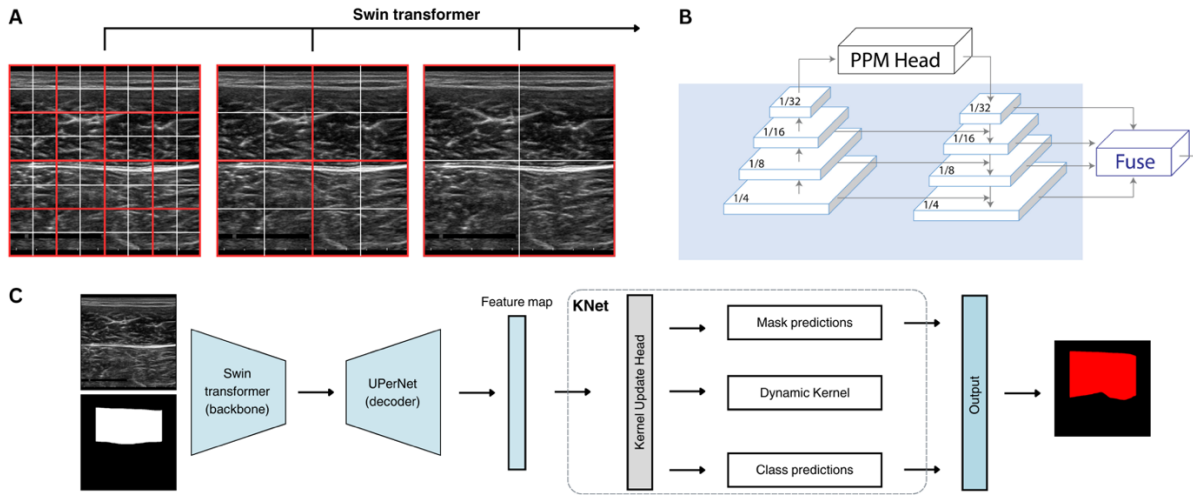


Figure 20 Segmentation framework used in this work. (A) Features are extracted by using a Swin transformer as backbone. (B) Multiscale features aggregation is performed by the UPerNet (Unified Perceptual Parsing for Scene Understanding). (C) Overall architecture of the Knet. During the inference process, the network generates a softmax output that indicates the probability of each pixel belonging to one of the muscle or background classes.

5.5. Pipeline development and evaluation

For the training and evaluation of the segmentation network, the dataset was divided into 5 different folds for cross-validation. As in previous works, the network was iteratively trained on three folds while one-fold was allocated for the validation set and one-fold was held out for testing purposes.

During the training phase images were scaled in the [0,1] range fully utilizing the uint8 dynamic, hence standardizing the images with a mean of 127.5 and a standard deviation of 127.5. To avoid overfitting standard on-the-fly augmentations were implemented by cropping, flipping, and using photometric distortion on the images.

Following the developer's implementation, for each run, the total number of iterations was set to 80000, with AdamW optimizer and MultiStep Learning Rate scheduling. The batch size to 2. For the loss function, CrossEntropyLoss loss was used with different weights for each class. Since the number of background pixels is way larger than the pixels of each muscle class, during loss calculation, weights are set for each class in a manner inversely proportional to their frequency in the dataset. A logarithmic transformation is applied to these inverse frequency weights to help in smoothing the distribution of weights, thereby mitigating the impact of extreme variations in class frequencies.

The inverse frequency weight for each class *c* belonging to the set *C* is then calculated as:

$$w_c = \frac{N}{N_c} \tag{1}$$

Where N is the total number of pixels and N_c the number of pixels belonging to that class. To normalize these weights, they are scaled by the minimum weight:

$$w'_c = \frac{w_c}{\min_{c \in C} w_c} \tag{2}$$

Finally, the logarithmic weights are obtained by applying the natural logarithm to the normalized inverse frequency weights, with an addition of 1 to avoid logarithm of zero:

$$w''_c = \log(1 + w'_c) \quad (3)$$

During training, the mean Intersection over Union was tracked to save the best model configuration and to assess the performance of the model on the validation set. The trajectory of IoU in the validation during training for all the runs is available in Figure 21.

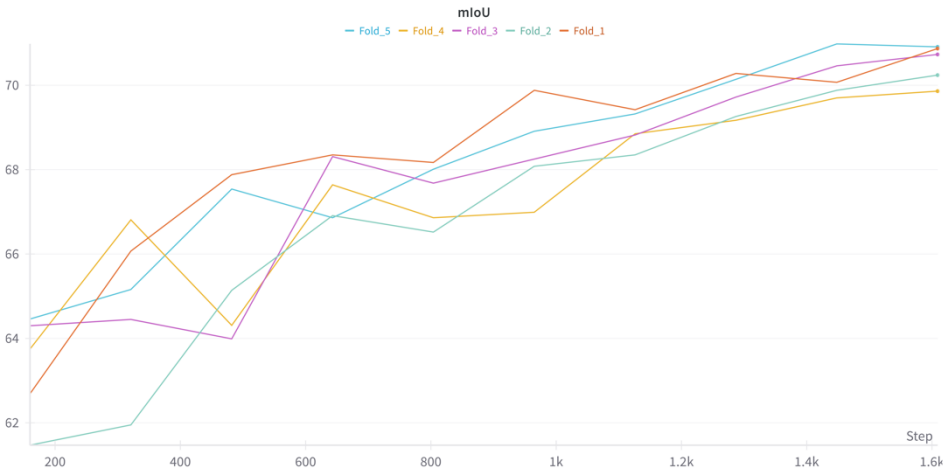


Figure 21 Mean Intersection over Union for different runs on the validation set.

During inference phase the image is standardized as in the training phase. For each class a segmentation map is available after the application of the Sigmoid operator on network output. At this point a post-processing is applied. It involves three primary steps: first, on the convex hull of all detected object is extracted. Overlapping objects of different classes (IoU > 0.6) are merged based on the label covering the biggest area. Second, for each label, the algorithm focuses on the largest connected component, discarding smaller, spurious segments. Finally, the algorithm determines the dominant label in the refined prediction, assigning a general class prediction based on the most represented label.

The final output of the network is therefore the classification of the image based on the most represented label, and the segmentation maps of one or more muscles.

After the segmentation step, to obtain information on the muscle in the image, texture features have been extracted using PyRadiomics [203]. The extraction has been customized focusing the extraction on the CSA using the segmentation map and extracting the 93 features of the following groups: first-order, GLCM, Gray Level Run Length Matrix (GLRLM) [181], Gray Level Size Zone Matrix (GLSZM) [204], Gray Level Dependence Matrix (GLDM) [205], and Neighbouring Gray Tone Difference Matrix (NGTDM) [206]. During manual Heckmatt scoring, the physician looks also for visual clues outside CSA like presence of bone or other structures. For this reason, the area below the muscle is also considered and a separate feature extraction is performed in this area. A visualization of this process is available in Figure 22 with the CSA of the muscle in green and the lower part in red (Figure 22, panel B). The lower part is defined in its top limit by the centroid of the CSA lowered by 10% of image height and by the CSA, while the lateral and lower limits are defined by the 15% of image width and height.

A classifier was trained to classify the images in three classes derived from Heckmatt score: Normal (Heckmatt 1), Uncertain (Heckmatt 2), Abnormal (Heckmatt 3+4). Heckmatt 3 and 4 were merged

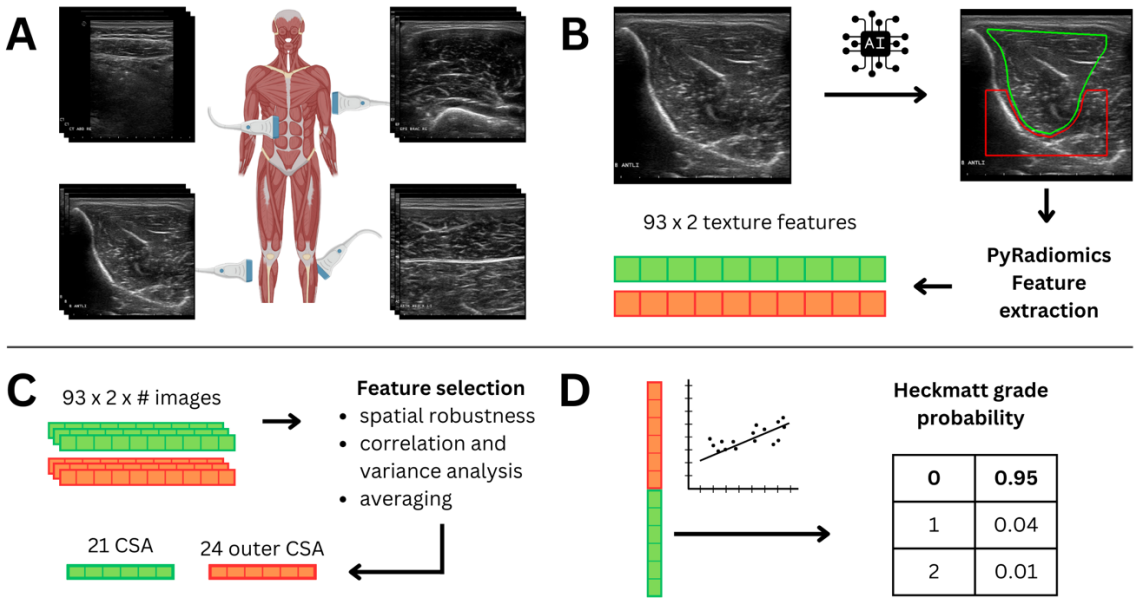


Figure 22 Project pipeline. A) Multiple repeated bilateral acquisitions. In the figure are shown examples from Rectus Abdominis, Biceps Brachii, Tibialis Anterior, Gastrocnemius Medialis B) The segmentation network performs muscle classification and segmentation, features are extracted from the CSA and from the area under the muscle using PyRadiomics. C) Feature dimensionality is reduced considering spatial robustness and correlation. Remaining features are averaged if multiple acquisitions are present. D) The ordinal regression model predicts probability of Heckmatt class.

given they bring similar information in terms of clinical evaluation. Features were measured for each subject-muscle-side combination for a total of 1466 entries and checked for spatial robustness and collinearity. To avoid error propagation the classifier was trained on features extracted from ground truth segmentation. Having more than one acquisition for each image, to calculate a specific feature for that image an average was performed across different acquisitions. If the range of a feature on the multiple acquisitions was more than 25% of its average value that feature was marked as "high spatial variability" for that image. Only features with a "high spatial variability" for less than 10% of the dataset (146) were retained in the model. Then features were standardized and checked for collinearity and if two features had a Pearson's correlation of more than 95% only the one with the highest variability was retained, resulting in 21 features for the CSA and 24 for the area below the muscle.On this set of features, Heckmatt scoring was modeled as an ordinal classification task using Python mord library [207] implementing an ordinal logistic model (All-Threshold variant). The classifier was trained using 10-fold cross validation stratified by Heckmatt score and grouped by subject. Feature importance was estimated by the weights of the logistic model and averaged between folds.

This research has three main objectives: the correct classification of the main muscle in the image, its segmentation, and the correct description of the texture of the muscles. For assessing the classification performances, the confusion matrix and its descriptors were computed. The segmentation was assessed using Intersection over Union as in the training phase, moreover Precision and Recall were measured for understanding the tendency of the model to under- or over- segment the images. Finally, the features extracted from both the manual and automatic masks were compared using Pearson's correlation.

All the results will be reported on the whole dataset as the union of the prediction of the five different test folds.

5.6. Results

The classification report for the classification task is available in Table 11 and the normalized confusion matrix in Figure 24. For all the muscles, the networks show impressive performance with the exception the Depressor anguli oris and the Orbicularis oris that are often confused by the network. Across folds the weighted average ranges from 0.963 to 0.968 showing great robustness to dataset sampling. Examples of segmentation results compared to ground truth are available in Figure 23.

The segmentation results display different performance across the different muscle types. From Table 12 it's clear how for bigger muscles that usually present well defined aponeuroses segmentation accuracy is higher than the one for facial muscles that are often indistinguishable from the surrounding structures. From the analysis of precision and recall it's evident how the network has a tendency to over-segment the CSA.

From Figure 25, the robustness of segmentation performances, measured by mean Intersection over Union, across folds can be appreciated. Again, both the body of the boxplots and the presence of outliers are in line with each muscle's difficulty level as discussed in the next section. The correlation analysis on the texture features showed that 22 out of 93 features had a correlation above 0.95, 44 above 0.90, and 81 above 0.8.

Heckmatt grading performance are reported in Figure 26. It is clearly noticeable how the uncertainty of the model lies in the boundaries between consecutive classes with only one sample confused from Normal to Abnormal. Similarly, ROC curves show very robust performances in evaluating the extreme classes and a slightly better accuracy in discerning class Uncertain to Abnormal than Normal to Uncertain. From Table 13 it is clear how adding features from the lower part of the image improves all the classification metrics compared to only using one of the two groups.

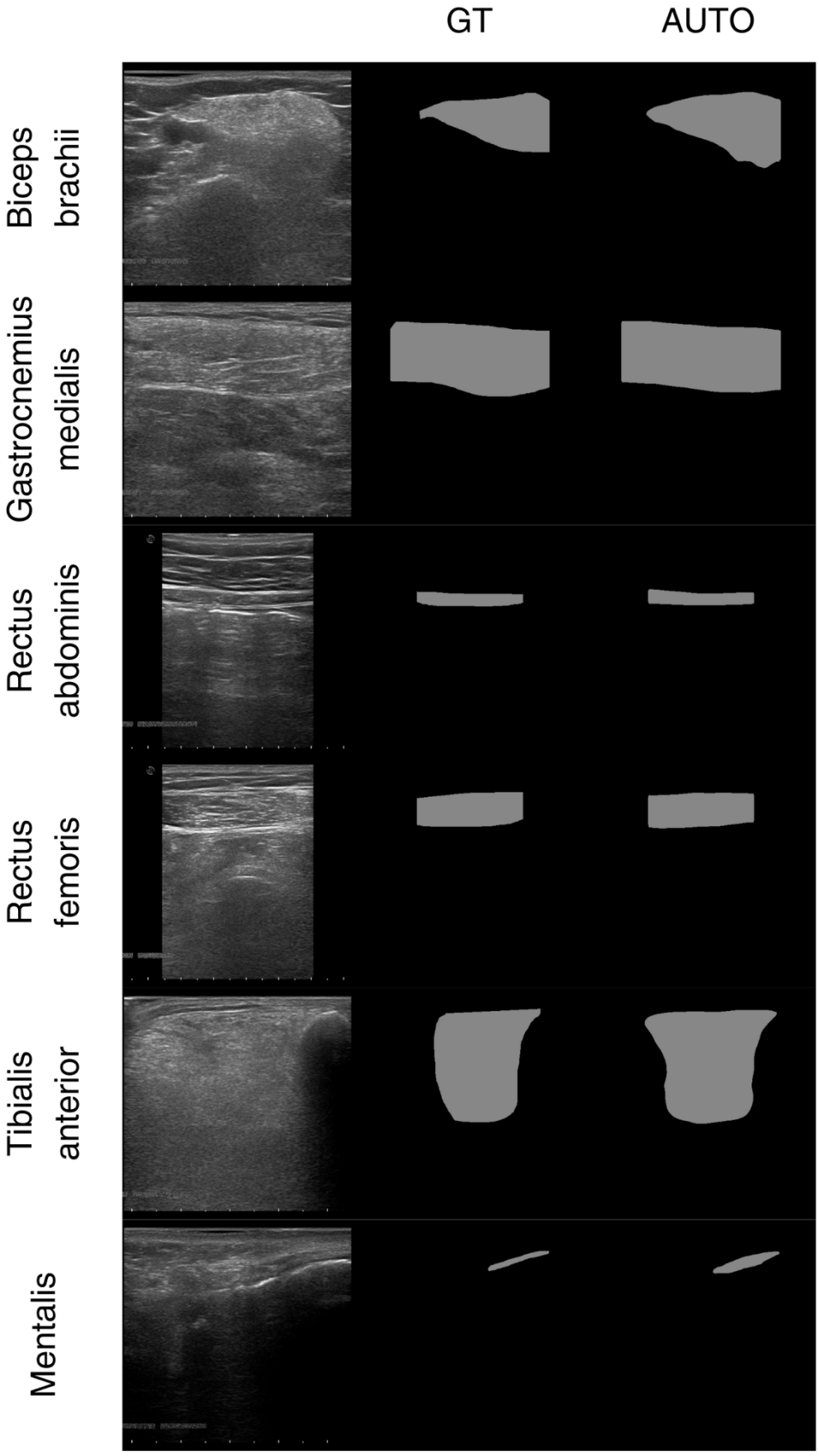


Figure 23 Segmentation results for different muscle types for the same subject.

Muscle	precision	recall	f1-score	support
Digastricus	0.992	0.983	0.987	1440
Masseter	0.998	0.997	0.998	1478
Depressor anguli oris	0.720	0.777	0.747	1258
Geniohyoideus	0.975	0.986	0.980	738
Temporalis	0.998	0.998	0.998	1471
Gastrocnemius medial head	0.999	0.997	0.998	1747
Orbicularis oris	0.814	0.753	0.783	1584
Vastus lateralis	0.981	0.986	0.983	1725
Mentalis	0.988	0.982	0.985	1243
Rectus femoris	0.981	0.979	0.980	1737
Zygomaticus	0.996	0.995	0.996	2571
Rectus abdominis	0.997	0.990	0.993	1732
Biceps brachii	0.997	0.996	0.997	1735
Deltoideus	0.990	0.998	0.994	1029
Tibialis anterior	0.997	0.999	0.998	1736
Trapezius	0.997	0.999	0.998	1522
accuracy	0.966	0.966	0.966	0.965
macro avg	0.907	0.907	0.907	24746
weighted avg	0.967	0.966	0.966	24746

Table 11 Classification report for muscle assessment.

Muscle	IoU	Precision	Recall
Biceps brachii	0.84 +/- 0.11	0.86 +/- 0.11	0.97 +/- 0.07
Deltoideus	0.86 +/- 0.08	0.87 +/- 0.08	0.98 +/- 0.03
Depressor anguli oris	0.47 +/- 0.28	0.49 +/- 0.30	0.73 +/- 0.41
Digastricus	0.73 +/- 0.13	0.75 +/- 0.13	0.96 +/- 0.13
Gastrocnemius medial head	0.90 +/- 0.08	0.91 +/- 0.07	0.99 +/- 0.05
Geniohyoideus	0.71 +/- 0.14	0.75 +/- 0.14	0.94 +/- 0.15
Masseter	0.84 +/- 0.08	0.85 +/- 0.09	0.98 +/- 0.05
Mentalis	0.41 +/- 0.16	0.44 +/- 0.17	0.85 +/- 0.24
Orbicularis oris	0.33 +/- 0.27	0.42 +/- 0.33	0.55 +/- 0.42
Rectus abdominis	0.81 +/- 0.10	0.83 +/- 0.11	0.98 +/- 0.06
Rectus femoris	0.85 +/- 0.10	0.86 +/- 0.10	0.99 +/- 0.04
Temporalis	0.74 +/- 0.20	0.78 +/- 0.19	0.95 +/- 0.13
Tibialis anterior	0.75 +/- 0.14	0.79 +/- 0.15	0.96 +/- 0.07
Trapezius	0.78 +/- 0.13	0.81 +/- 0.13	0.96 +/- 0.08
Vastus lateralis	0.86 +/- 0.11	0.88 +/- 0.10	0.98 +/- 0.07
Zygomaticus	0.39 +/- 0.25	0.42 +/- 0.27	0.69 +/- 0.38

Table 12 Segmentation performances across different muscle types.

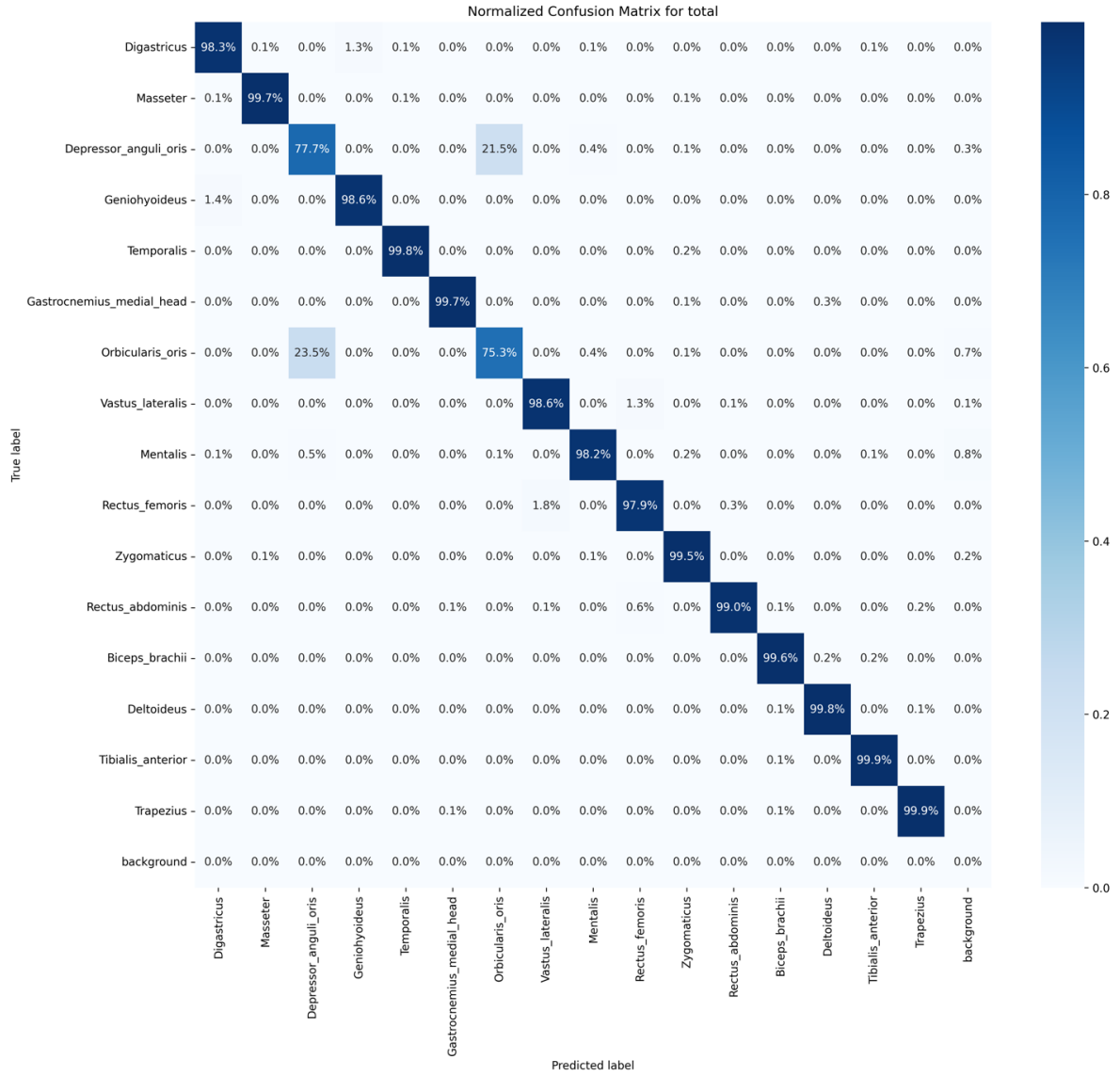


Figure 24 Normalized confusion matrix (whole dataset N=24756).

	# features	Accuracy	Kappa	AUC (One-vs-One)			AUC (One-vs-Rest)		
				01	02	12	0	1	2
CSA + Outer	45 (21)	0.80	0.70	0.92	1.00	0.89	0.96	0.87	0.97
CSA	21	0.78	0.67	0.91	1.00	0.87	0.95	0.86	0.97
Outer	24	0.74	0.60	0.86	0.99	0.83	0.92	0.80	0.96

Table 13 Ablation Study on feature extraction location. Number of features refers to the features selected during the fitting of the Heckmatt regression model.

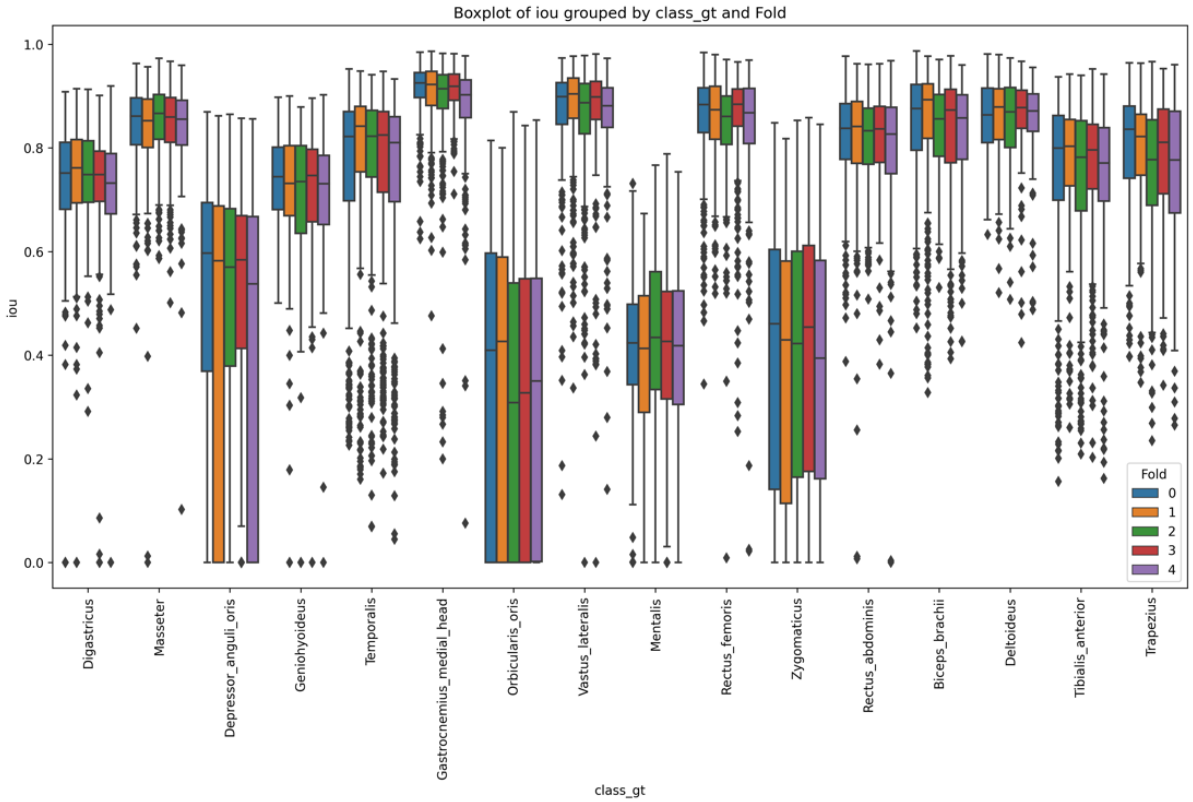


Figure 25 Boxplots for IoU by fold.

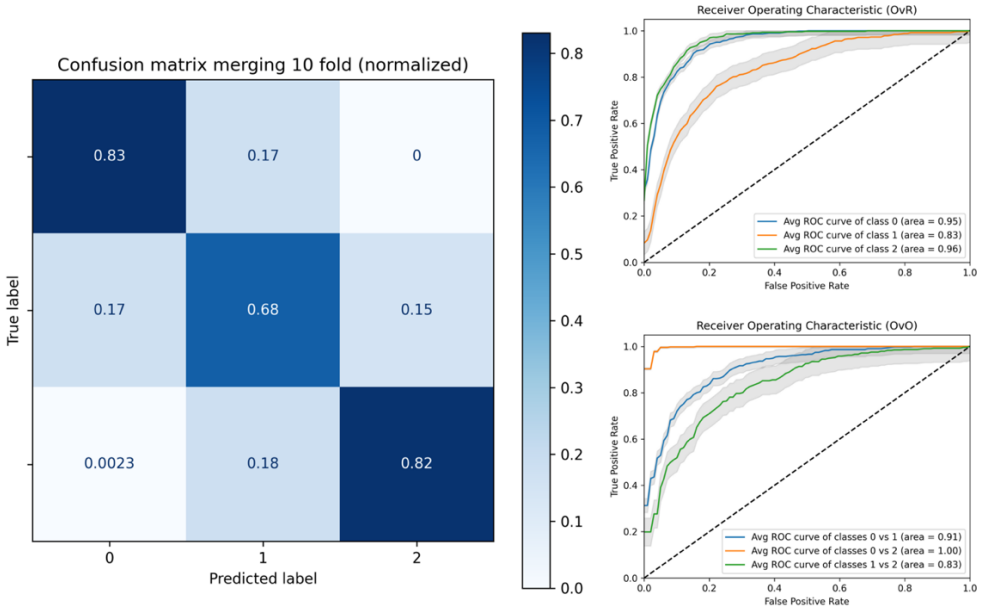


Figure 26 A) Normalized confusion matrix for prediction of Heckmatt grade. B) (top right) ROC One-vs-Rest approach. C) (bottom right) ROC One-vs-One approach.

5.7. Discussion

The field of muscle ultrasound (MUS) presents several challenges, especially in the context of deep learning applications. One significant hurdle is the inherent complexity of muscle tissues, which exhibit heterogeneous textures and often indistinct boundaries. This necessitates highly adaptable and precise segmentation algorithms. The availability of open datasets can significantly enhance the development of deep learning models for MUS. Open datasets allow for more robust and diverse training, leading to models that are better generalized and can perform well across different settings and equipment. This is crucial given the variability in ultrasound machine settings and patient-specific factors like age and obesity, which can influence ultrasound results. Therefore, ensuring the correctness of information extracted from images is paramount, as it underpins the reliability of diagnostic and therapeutic decisions. A standardized acquisition protocol for muscle ultrasound is vital. Consistency in image acquisition, such as maintaining uniform settings across different ultrasound measurements, ensures that the resultant data is more uniform and comparable. This uniformity is particularly important for developing and training deep learning models, as it reduces the variability in the data that the models need to account for.

A fully automatic pipeline capable of adapting to different muscle types was presented in this chapter. Such a pipeline should be able to handle the wide range of muscle textures, sizes, and shapes encountered in practice. The tendency of networks to over-segment, shown in Table 4, can be caused by manual annotations that tend to be conservative due to a focus on muscle texture. This tendency is also exacerbated by the pooling layers in neural networks, which can lead to the loss of fine details and result in over-segmentation. This is not a problem when comparing texture features for most of the descriptors calculated in this context, but it's a limitation that has to be taken into consideration if the goal is to extract a clinical descriptor or a diagnosis.

Regarding the segmentation of facial muscles, the task is more challenging compared to larger muscles. This is evident from the lower segmentation accuracy for facial muscles observed in the results. The smaller size and the entangled structures of facial muscles, often indistinguishable from surrounding tissues, make them more difficult to classify and segment accurately.

The results related to the Heckmatt classification task are in line with the clinical practice where the most difficult part for the algorithm and the physician is to correctly classify between neighboring classes. The choice of using PyRadiomics features and avoid having a fully DL based pipeline was taken to extract from the model the information about the most important features valued by the model, with the objective of mining new insights on Heckmatt scoring. The most important features for the model are related to the concepts of Entropy and Non-Uniformity in GLRLM and GLDM matrices, moreover the average gray value is the third most important features. This higher order features are linked to the presence of highly echogenic and disorganized structures inside the CSA, that are often caused by fatty infiltration, while a high echogenicity could be also due to strong reflectors such as the medial aponeuroses in the Tibialis anterior. This correspondence between visual interpretation of the images and their mathematical description is key for developing an AI system that could be of support in clinical decisions.

For future research one of the most interesting steps would be the integration of the segmentation and classification pipeline with a real-time system. This system would incorporate temporal and patient-specific information, thereby enabling a multimodal approach to muscle ultrasound analysis. Such an approach would leverage the strengths of various data types, potentially leading to more accurate and holistic assessments of muscle health.

In conclusion, the application of deep learning to muscle ultrasound imaging holds great promise but is not without its challenges. Addressing these challenges through developing robust, adaptable models and integrating multimodal data will be crucial for advancing this field.

6. A modular approach for medical image simulation uncoupling semantic and stylistic information in a limited data environment

6.1. Introduction

From the previous chapters, it is clear how in the medical image domain, and especially in ultrasound image analysis, having a correct representation of both the structure and the texture of the underlying morphology is of great importance to be able to extract useful clinical information from the image data. Limitations of discriminative models in understanding image data could originate from different sources. Under representation of specific subgroups in the datasets or uncertainty in the manual annotations lowers the precision of models trained on those data. Moreover, suboptimal choices in loss function could affect the representation of the underlying task.

In the chapter about Intima Media Complex (IMC) segmentation the segmentation accuracy and therefore the accuracy of downstream analysis was lowered by the variability in the manual tracings of the IMC boundaries. Optic Nerve Diameter (OND) measurement was challenged by the low SNR of the images and the scarce numerosity and high variability of the dataset available. In the automatic muscle analysis, facial muscles presented a challenge often too hard for the segmentation network.

Some of these challenges could be addressed by using generative models. By producing high quality medical images, generative models can improve the way new deep learning algorithms are implemented and used. In the tasks discussed in previous chapters, generative models could be used to improve image quality, either by denoising, image standardization, or image enhancement. While dealing with facial muscles in muscle analysis, augmentation could be a way to improve the training set for segmentation network training.

Aside from only improving existing images and datasets, having control over the generation enables using the generated images both for validating discriminative models and, most importantly, to create teaching tools for the training of physicians. A high-quality and realistic dataset generated from a known ground truth could be used as benchmark for discriminative models. Having the possibility to generate different level of pathology or rare diagnostic condition could be very helpful in training physicians that otherwise might encounter similar examples only few times in their careers.In recent years the improvement in image synthesis has been tremendous thanks to transformers and diffusion models [93], [208], [209]. While diffusion models reach unprecedented levels of image quality and model flexibility, they require an iterative process during inference incurring high computational cost in contrast with GAN that generates an image with a single pass.

To achieve controllable generation, a popular line of research focuses on generating and editing images conditioned on certain guidance. Typically, visual clues, such as segmentation maps and sketch maps, have been widely adopted to guide image synthesis and editing [79], [210]. More recently text-guided diffusion models achieved impressive results in conditional image generation, but they require huge datasets and can achieve fine detail control only through editing [211]. Another approach to control generative models stems from conditioning the generation process through auxiliary network but still requires datasets in the order of thousands of images [212].

In this chapter a modular pipeline is proposed to generate ultrasound images in a controlled way even with less than 1000 samples available. The same concept will then be applied in the field of digital

pathology demonstrating the adaptability off the proposed approach. The generation is split into three main parts: semantic content generation, image generation, and style generation. The pipeline is modular in the sense that for each component the inputs can come from different sources, such as real data, human defined data, or AI generated while the general approach would be still applicable.

Following the schematics in Figure 27 the first step is generating the semantic content. This can be done by an AI model, a physician that draws a particular structure or this info can be retrieved from real data using manual or automatic methods. This step controls the level of semantic detail in the generation process. From this point the semantic map is available. The next step is the generation of ultrasound images. If the task is to generate an image with a known morphology or other parameters to condition the generation process, two roads are available. The first is to use a generative model to go directly from the segmentation map domain to the image domain. The other one is to generate the image using a physics-based simulation. This involves commonly two steps: the first is to estimate the physical properties of each tissue and the second is to simulate the interaction of sound waves with the underlying structure.

The last part of the pipeline involves the generation or control of the stylistic content of the image, hence controlling the appearance of the image without changing its semantic content. This step can potentially estimate from the same image how it can appear changing acquisition settings without changing the semantic content of the image. This step could be applied either to real image as to images generated by the previous steps.

This chapter is structured in the following manner: Initially, it presents an introduction to similar studies that have explored ultrasound image simulation through generative models. Subsequently, it introduces the application in the field of thyroid ultrasound imaging. After that, the application proposed in this study is elaborated upon, detailing both the dataset and the model employed. This is followed by an exposition of how the approach was validated. The chapter then extends the discussion to the application of the proposed pipeline within digital pathology, adhering to the same organizational structure. The chapter concludes with a section dedicated to a general discussion.

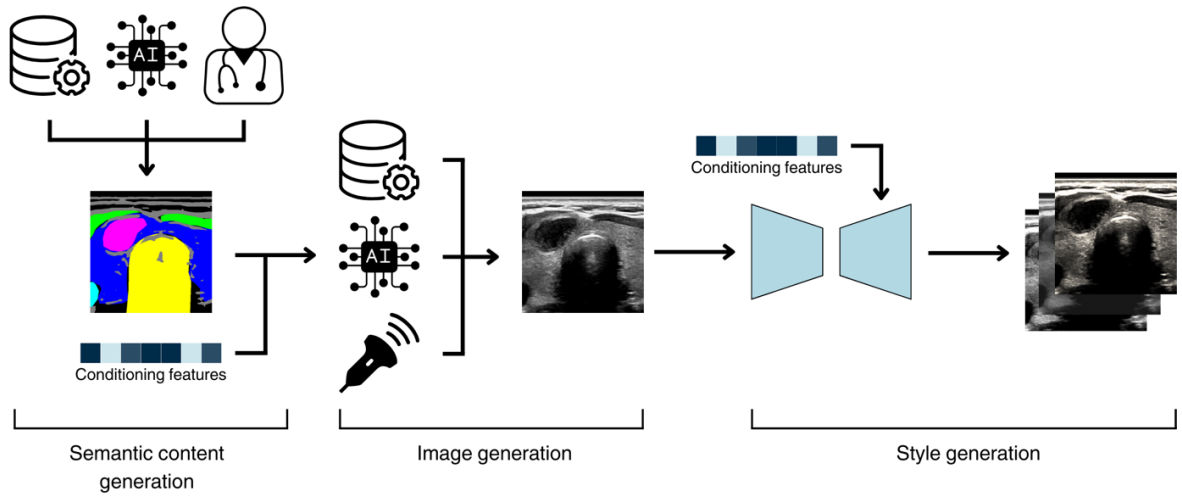


Figure 27 Pipeline for controllable image generation.

6.2. Related work

The development of generative models for ultrasound image synthesis has significantly advanced, addressing challenges such as data scarcity, annotation requirements, and the need for high-quality, diverse image datasets. Recent works have demonstrated the efficacy of various generative adversarial networks (GANs) and diffusion models in synthesizing realistic ultrasound images.

Deep convolutional generative adversarial networks (DCGANs) [75] have been effectively employed for generating breast ultrasound images, highlighting their potential in creating realistic, clinically valuable synthesized images. Fujioka et al. [213] have shown that DCGAN can produce high-quality images that are indistinguishable from original scans, emphasizing the model's utility in enhancing the quality of medical imaging datasets without the need for extensive data collection. Following this work they applied DCGANs for virtual interpolation of tumor development in breast ultrasound further underscoring the potential of generative models to aid in visualizing tumor progression, offering a novel tool for educational and research purposes [214]. CycleGANs were used in [215] to generate realistic B-mode musculoskeletal ultrasound images again to make large, labeled datasets for training image segmentation networks. The experiments resulted in realistic images but with reduced diversity.

In [216] Liang et al. use sketch-guided progressive growing GANs (PGSgan) for synthesizing editable and high-resolution ultrasound images. This approach addresses the challenges of high-resolution image synthesis and offers a mechanism for generating images with fine-grained structural details. CycleGANs have been utilized to generate synthetic labeled data from existing anatomical models, demonstrating the capability of GANs to produce realistic ultrasound images alongside paired labels [217]. This approach significantly reduces the effort required for manual annotation, facilitating the generation of extensive datasets for training deep learning models.

A 3D implementation of the Pix2Pix architecture has been explored for the synthesis of 3D cardiovascular ultrasound images. This development is crucial for overcoming the scarcity of 3D labeled datasets, enabling the advancement of deep learning applications in 3D medical imaging [218]. To overcome some of GANs limitations and improve diversity of generated images, adversarial denoising diffusion models (DDMs) have been introduced for generating echocardiography images. This method enhances the efficiency of generating new data, offering a robust framework for clinical research [219].

DDPMs have also been applied directly with a semantic label input to generate synthetic 2D echocardiograms to train a segmentation network with, yielding a mean Dice score of more than 80% for left ventricular endocardium, epicardium and left atrial segmentation respectively [220].

In this chapter a semantic image synthesis that generates both physiological and pathological conditions is explored. It employs a pyramidal GAN that can generate a realistic thyroid ultrasound image correctly representing the texture features of different tissues.

In conclusion, the scope of research in ultrasound image synthesis using generative models, from DCGANs to diffusion models, underscores the significant potential of these techniques in medical imaging. The advancements in controllable image generation, particularly in the synthesis of high-resolution, anatomically accurate images, present an asset for both medical education and the development of automated diagnostic tools.

6.3. Thyroid ultrasound

Ultrasound imaging plays a key role in the evaluation and management of thyroid conditions, given its non-invasive, real-time, and cost-effective nature. It is particularly effective in assessing thyroid nodules, which are prevalent in clinical practice, affecting over 50% of the population[221]. These nodules, varying in etiology, can be benign or carry a risk of thyroid cancer, underscoring the need for accurate diagnosis [222].

Ultrasound parameters such as microcalcifications, hypoechogenicity, and nodule shape are traditionally linked to malignancy risk, but none are reliable in isolation [223]. Advanced techniques like elastography and fine-needle aspiration (FNA) biopsy augment the diagnostic process, yet they have limitations.

Thyroid ultrasound imaging is essential for several reasons:

- **Anatomical Assessment**: It helps in evaluating the symmetry of thyroid lobes, understanding the neck's anatomical orientation, and examining relationships with surrounding structures.
- **Detailed Examination**: Longitudinal scans are crucial for analyzing the homogeneity of thyroid tissue and identifying small pathological changes.

However, challenges exist in ultrasound-based diagnosis:

- **Subjectivity and Variability**: The interpretation of ultrasound images depends heavily on the experience of the operator, leading to inter- and intra-observer variability.
- **Image Quality**: The quality of image acquisition significantly impacts diagnostic accuracy.
- **Nodule Classification**: Systems like the Thyroid Imaging Reporting and Data System (TIRADS) categorize nodules based on malignancy likelihood, but the classification can be complex.

In response, computer-aided diagnosis (CAD) techniques have been developed to assist in detecting, segmenting, and classifying thyroid nodules. Deep learning models have shown higher accuracy compared to traditional methods. However, challenges such as the need for large, high-quality datasets and artificial markers in images, which can misguide the models, persist.

6.4. Dataset description

The Montgomery Dataset, a collection of 735 thyroid B-mode ultrasound images, has been manually segmented by an expert, assigning specific labels to various semantic features. The labels identified in the image are the following:

- LABEL 0 → internal background
- LABEL 1 → connective tissue
- LABEL 2 → trachea
- LABEL 3 → thyroid gland
- LABEL 4 → blood vessels
- LABEL 5 → dermal
- LABEL 6 → muscle tissue
- LABEL 7 → thyroid nodule
- LABEL 8 → external background

The images vary in size, including dimensions from 1024x768 to 960x720, but were uniformly cropped to a 768x768 resolution to focus on the region of interest (ROI) and minimize the presence of the external background (Label 8).

To anonymize the images and ensure the model would not waste capacity in learning ancillary information, non-essential information such as patient data, machine specifications, and acquisition parameters were removed. This was achieved by adjusting all pixels associated with the external background label to black (value 0), as depicted in Figure 28. Panel A in this figure illustrates the original images with auxiliary information (already anonymized), Panel B shows the adjusted images with the external background removed, and panel C shows the final patch extraction process. To select the ROI for patch extraction, the centroid of the complementary of the external background label was selected as the patch center using a 768x768 patch.

To train and validate the model, images were resized to 512x512 using bicubic interpolation for images and nearest interpolation for segmentation masks due to limitation in GPU memory for training the deep learning model. Moreover, training the model at full resolution but with limited capacity was also tested but resulted in lack of convergence by the network.

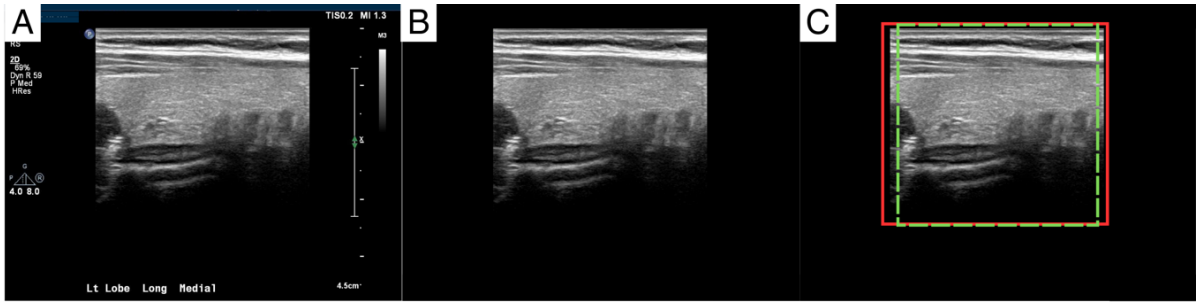


Figure 28 Process for determining best crop for ultrasound image.

6.5. Model development

The chosen model for semantic image synthesis in this study is the Dual Pyramid Generative Adversarial Network (DP-GAN). This innovative architecture excels in learning and applying spatially adaptive normalization across various scales. It uniquely leverages scale information in a bidirectional manner, unifying supervision at different levels.

Traditional semantic image synthesis methods, using encoder-decoder architectures with semantic label maps, often lose detail due to limitations in normalization techniques like instance normalization. DP-GAN addresses this by implementing spatially adaptive normalization, dispersing semantic information more effectively across the network, based on the input segmentation mask.

Previous discriminators in this domain, such as those used in SPADE [210] and OASIS [224], have limitations in producing high-fidelity images, particularly at varying object scales. DP-GAN overcomes these issues by employing multi-scale supervision, which includes multi-scale feature matching and adversarial supervision at the patch level. This approach ensures better alignment of generated images with input semantic maps, preserving essential scale details.

The dual pyramid structure of the DP-GAN generator is a key innovation. It comprises two pyramids: one for image synthesis and another for spatial adaptation. The generator takes as inputs a label map and a 3D noise tensor, enabling the creation of diverse images from the same map. A schematic representation of the DP-GAN generator is available in Figure 29. The DP-GAN discriminator goes beyond pixel-level loss, incorporating pixel, patch, and feature-level supervision for more realistic structure generation at all detail levels.

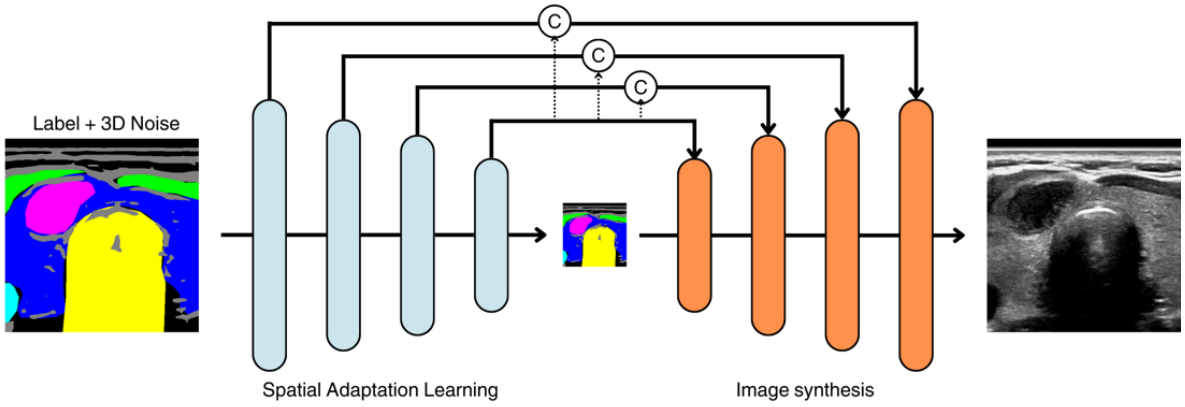


Figure 29 DP-GAN Generator. The first pyramid (orange) comprises spatially-adaptive normalization (SPADE) blocks and upsampling, which transforms a downsampled version of the input into an image. Meanwhile, the secondary pyramid (blue) is responsible for refining the conditioning of the SPADE blocks in the orange pyramid. Moreover, the lowest resolution layer output in the yellow pyramid undergoes upsampling and is then integrated (indicated by C) with the features from preceding layers.

The architecture of this network is designed to calculate the likelihoods for $N+1$ categories at every individual pixel, where N symbolizes the count of semantic categories, supplemented by an extra 'fake' category. In the phase of training, the authentic label for each pixel is established based on the label map of an actual image and is tagged with the 'fake' category for an image that is synthetically generated. To tackle the issue of class unbalance, a weighted cross-entropy loss function for $N+1$ classes is employed, defined as:

$$L_{pixel} = -E_{(\boldsymbol{x},\boldsymbol{l})}\left[\sum_{c=1}^{N}\alpha_c \sum_{x,y}^{H\times W} \boldsymbol{l}_{x,y,c}\log D(\boldsymbol{x})_{x,y,c}\right] + -E_{(\mathbf{z},\boldsymbol{l})}\left[\sum_{x,y}^{H\times W}\log D\big(G(\mathbf{z},\boldsymbol{l})\big)_{x,y,c=N+1}\right] \quad (1)$$

Here, $\boldsymbol{x}$ denotes the authentic image, $\boldsymbol{l}$ symbolizes the label map, and $(\mathbf{z},\boldsymbol{l})$ represents the 3D noise tensor with the label map. The term α_c signifies a distinct weight. To enhance the framework, two additional loss functions are incorporated, as relying solely on pixel loss proves inadequate. The first auxiliary loss function is a multi-scale adversarial loss based on patches, targeting feature maps at a reduced resolution:

$$L_{ms}^{i} = -E_{(\boldsymbol{x},\boldsymbol{l})}\left[\min\left(-1 + D_p^i\left(\psi^i(\boldsymbol{x})\right),0\right)\right] - E_{(\mathbf{z},\boldsymbol{l})}\left[\min\left(-1 + D_p^i\left(\psi^i\big(G(\mathbf{z},\boldsymbol{l})\big)\right),0\right)\right] \quad (2)$$

In this context, Equation 2, $\psi^i(\boldsymbol{x})$ and $\psi^i\big(G(\mathbf{z},\boldsymbol{l})\big)$ indicate the feature maps at the ith layer in the discriminator's encoder for either a real or a synthetically generated image. D_p^i is a convolutional network comprising blocks that include a convolution layer, ReLU activation, and batch normalization. This network yields predictions for each feature correlating to patches of diverse sizes within the original image. The loss L_{ms}^{i} is then averaged over all the layers.

The second supplementary loss is grounded in a multi-scale feature matching method that measures the L2 distance between the features of a real image and its synthetic counterpart. The key aim of this

loss is to motivate the generator to create images that are visually akin to their real equivalents in the feature space of the discriminator's decoder.

$$L_{fm}^{i} = -E_{(\boldsymbol{x},\boldsymbol{l},\boldsymbol{z})}\left[\frac{\sum_{x,y}^{H^i \times W^i} \left\| \phi^i(\boldsymbol{x})_{x,y} - \phi^i\left(G(\boldsymbol{z},\boldsymbol{l})\right)_{x,y} \right\|_2^2}{C^i \times H^i \times W^i}\right] \tag{3}$$

Equation 3 illustrates the loss for a single layer i. Here, ϕ^i symbolizes the feature vector with dimensions C^i , and the resolution at that layer is depicted by $H^i \times W^i$. The loss L_{fm}^i is also averaged across all layers.

The loss for the generator is formulated as in Equation 4, where each component reflects supervision at the pixel, patch, and feature levels respectively.

$$\begin{aligned} L_G = -E_{(\boldsymbol{z},\boldsymbol{l})}\left[\sum_{c=1}^{N} \alpha_c \sum_{x,y}^{H\times W} \boldsymbol{l}_{x,y,c} \log D\left(G(\boldsymbol{z},\boldsymbol{l})\right)_{x,y,c}\right] \\ -\frac{1}{L}\sum_{i=1}^{L} \mathbb{E}_{(z,l)}\left[\min\left(-1 + D_p^i\left(\psi^i\left(G(\boldsymbol{z},\boldsymbol{l})\right)\right), 0\right)\right] + L_{fm} \end{aligned} \tag{4}$$

Within the DP-GAN model, we incorporate the LabelMix regularization as proposed in [224]:

$$L_{LM} = \left\| D_{logits}\left(LabelMix(\boldsymbol{x}, \hat{\boldsymbol{x}}, M)\right) - LabelMix\left(D_{logits}(\boldsymbol{x}), D_{logits}(\hat{\boldsymbol{x}}), M\right) \right\|_2^2 \tag{5}$$

In this equation, D stands for the logits prior to the ultimate softmax layer, and M represents a binary segmentation mask utilized for merging real and synthetic images via LabelMix.

Consequently, the overall loss is articulated as:

$$L = L_G + L_{pixel} + L_{ms} + \lambda_{LM} L_{LM} \tag{6}$$

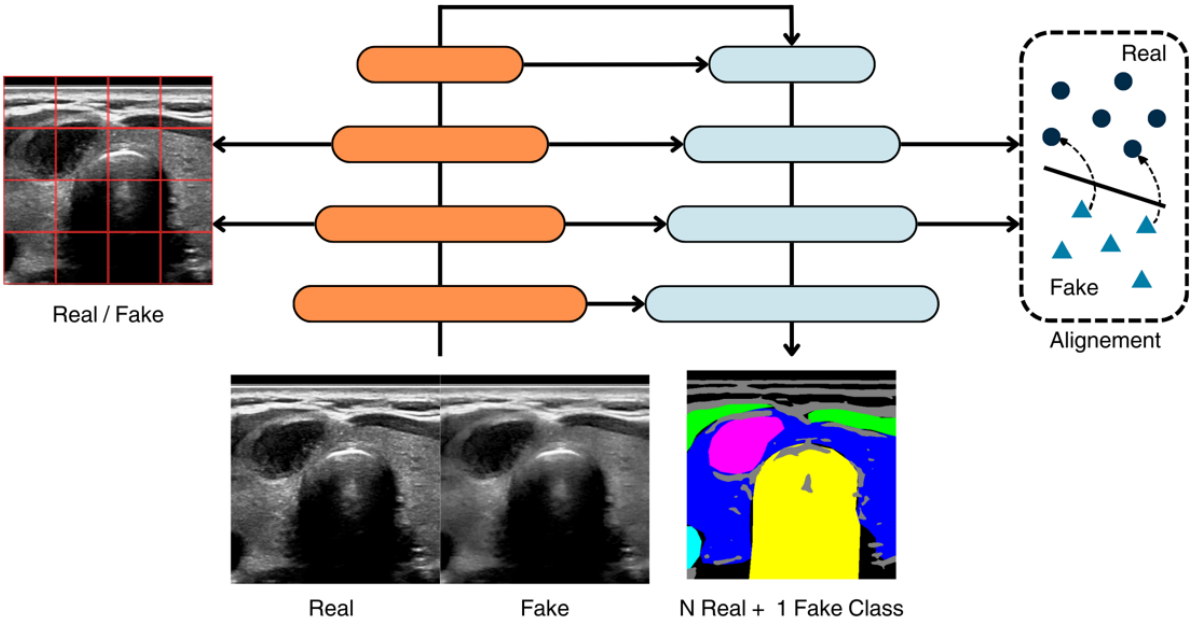


Figure 30 DP-GAN discriminator.

The network was defined using 6 ResNet blocks, 64 filters in the first convolution layer of the Generator, and a latent vector of dimension 64. The training was performed for 100 epochs with a batch size of 4 512x512 images. Adam was used as the optimizer with a learning rate of 0.0001 for the generator and 0.0004 for the discriminator. Checkpoint were taken every 5000 iteration both for single model and it's EMA version with a decay rate of 0.9999. FID metric [225] was tracked during training to select the best iteration. Losses and FID are visualized in Figure 31.

At each checkpoint inference was performed on the validation set and the SSIM and PSNR were computed between Real and Generated images to evaluate the quality of generated images. Since the aim is to represent each feature with a realistic appearance, we computed the texture features on each label with Pyradiomics and on the entire image excluding background. We compared the distributions of the features using LDA for the features relative to the specific labels and PCA to compare the features relative to the entire image. For each distribution comparison, we also compute Precision and Recall using a Pytorch implementation of the metrics originally available in Tensorflow [226], [227]. Precision and Recall were not computed for the thyroid nodules label due to a low prevalence in the dataset that caused the metric to go to 1.0 voiding the results. To have a task-based validation a model based on UNet trained to segment the different structures in the thyroid images was applied to the images generated at the different timepoints tracking the IoU.

6.6. Results

In Figure 31 it can be observed how discriminator losses between real and fake images gradually converge to similar values. This means that while the generator becomes more capable of making better images, the discriminator does not lose ability in discerning between real and fake image still providing a useful gradient for the network training. The FID path confirms this result. The quality of the generated images is high compared to the literature with a PSNR of 19.99 +/- 2.25 dB and an SSIM of 0.59 +/- 0.07.

As for the network fidelity in representing the correct textures in the image, from Figure 32 it's evident how during training the distribution of the features resulting from the fake images correctly shifts

towards matching the one obtained from real ones. As predictable using a GAN based approach, the precision in the feature space is higher than the recall even if they both show a similar path during learning. Focusing on each feature using a discriminant analysis like in Figure 34 it can be seen that for bigger structures with low echo intensity like the background or the trachea the network is able to produce a more reliable representation of those structures with respect to more varying structures like the thyroid gland or the connective tissue. Similarly, from the task-based validation in Figure 35 we can see different performances for the different label. This is both due to the inherent challenges of those structures already present in real data and in a more challenging representation of their features.

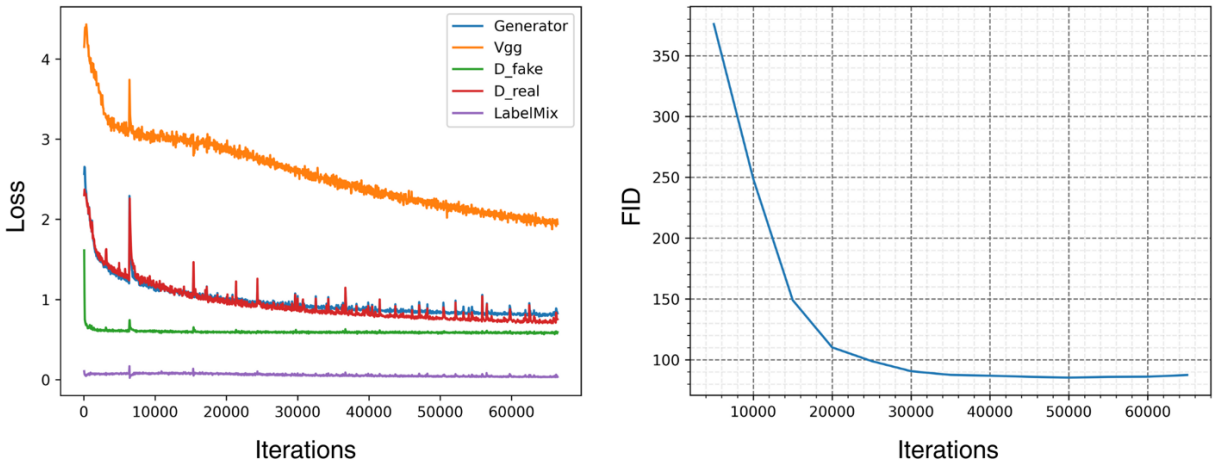


Figure 31 Losses and FID during training.

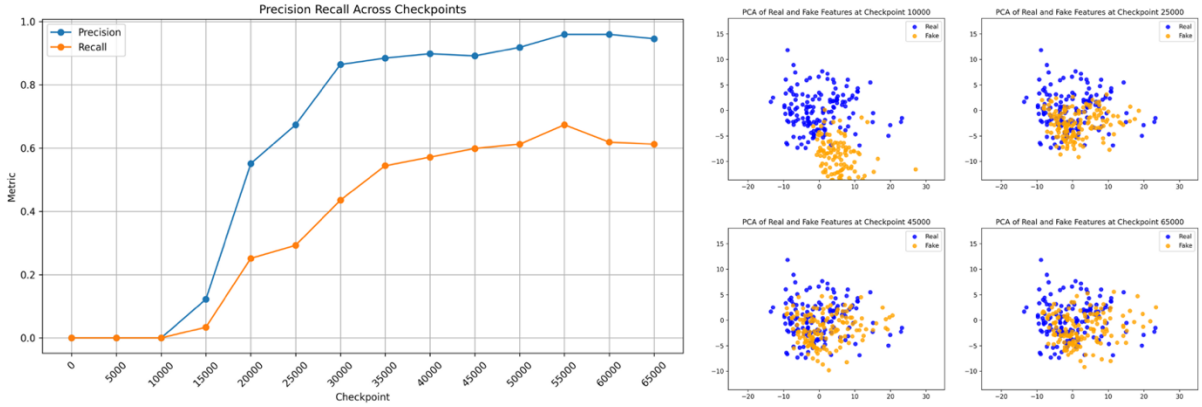


Figure 32 On the left Precision and Recall calculated using the radiomics features at the different timepoints. On the right, the plot of the two principal components of the PCA computed on the radiomics features in the image area at 4 different timepoints.

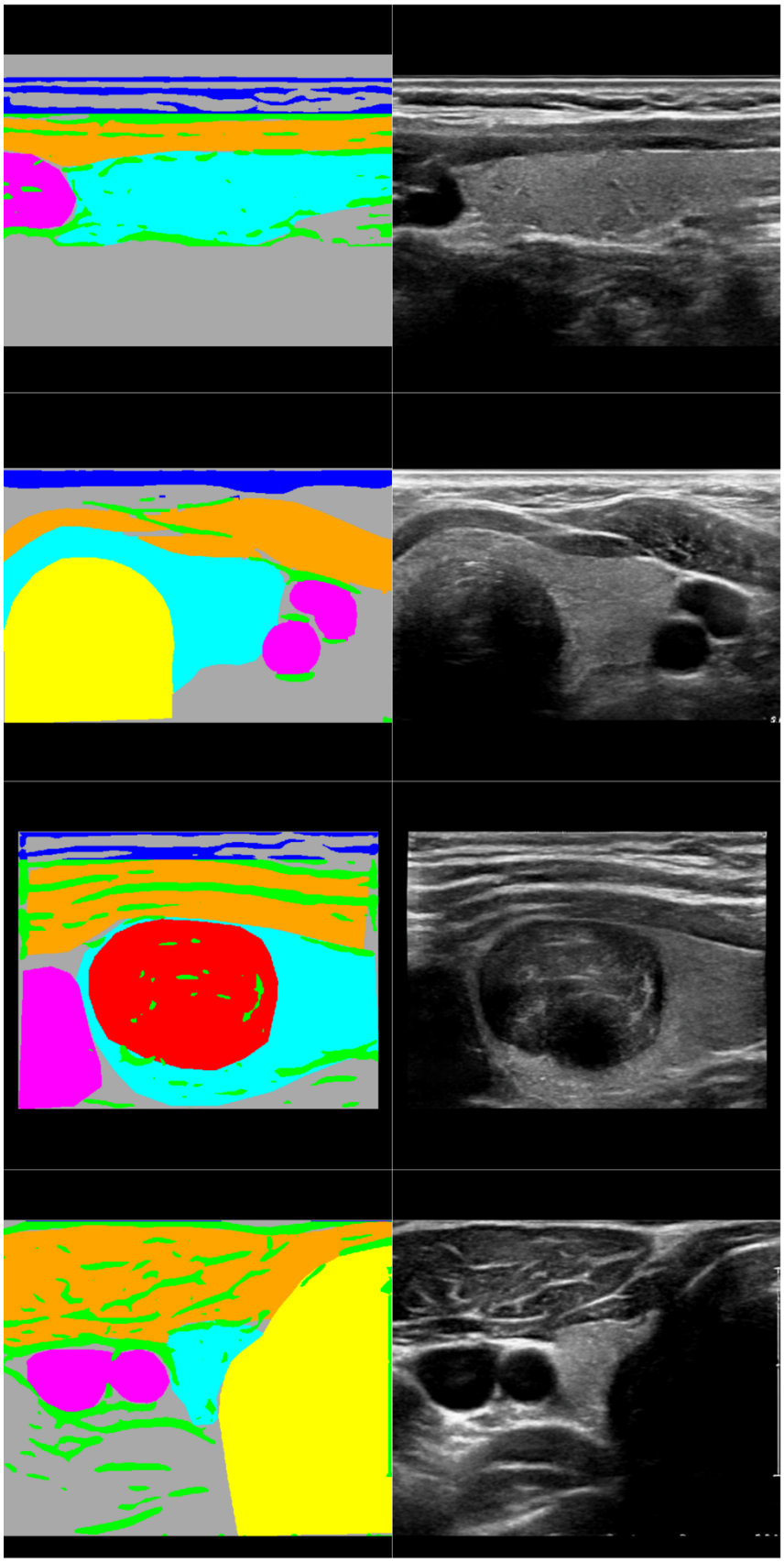


Figure 33 Examples of synthesized thyroid ultrasound images. In the segmentation map: black (external background), gray (internal background), yellow (trachea), cyan (thyroid), magenta (vessels), green (hyper echogenic structures), red (nodule), blue (skin and derma), orange (muscle)

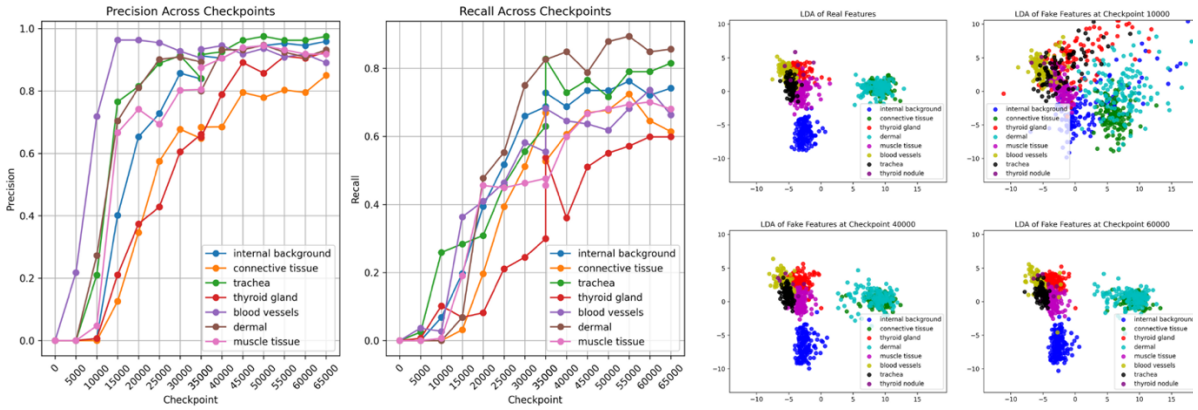


Figure 34 On the left Precision and Recall calculated using the radiomics features at the different timepoints for each label. On the right, the plot of the two principal components of the LDA computed on the radiomics features in each label at 4 different timepoints.

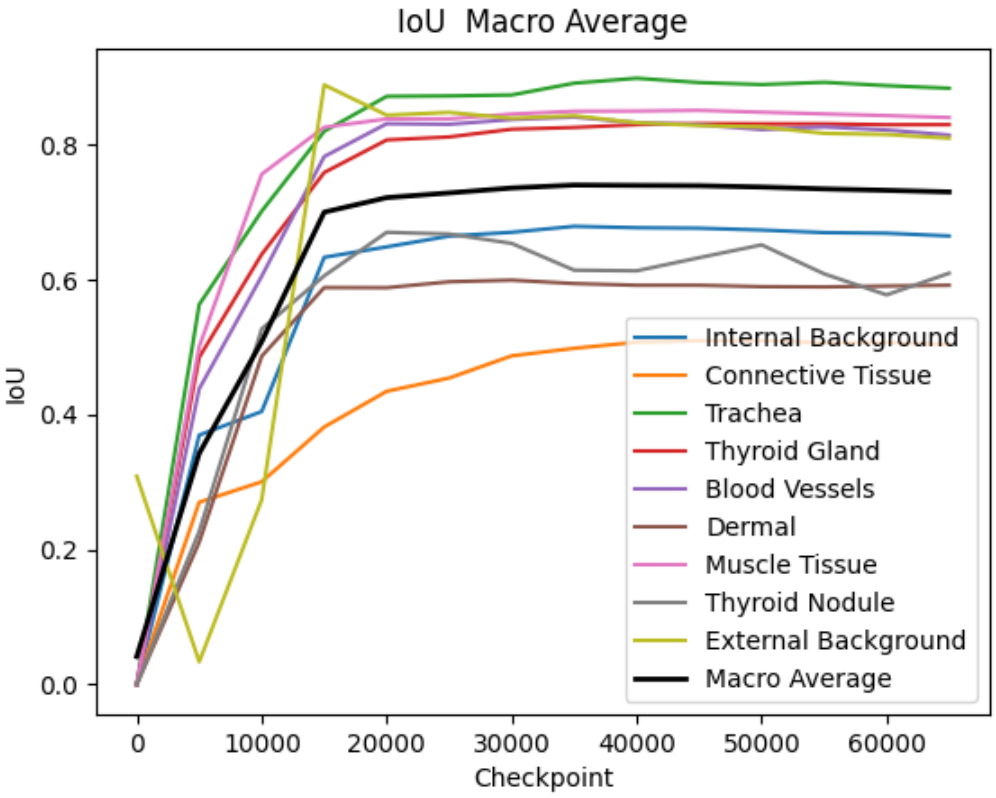


Figure 35 Intersection over Union of trained network at each checkpoint for each label.

6.7. Generative Modular pipeline for digital pathology

6.7.1. Clinical application

The applications of the proposed pipeline are not limited to ultrasound imaging. The problems of limited data affect all medical imaging modalities and for several diagnosis pipelines the semantic content is directly linked to the disease evaluation. One example that was investigated during the study of this pipeline is the simulation of digital pathology images of prostate tissue affected by adenocarcinoma.

Prostate adenocarcinoma is characterized by abnormal and heterogeneous glandular architecture [228]. The Gleason scoring system classifies this architecture by assigning a score between 1 and 5 to different histological patterns based on the extent of glandular differentiation and the tumor growth pattern within the prostate stroma. Pathologists evaluate the structural characteristics of the sample at low magnification (from 4x to 10x), following guidelines illustrated over the years by the International Society of Urological Pathology (ISUP) [229]. The Gleason groups are determined by the specific histological features observed in the tumor samples, as follows:

- Grade Group 1: Only single, well-differentiated glands.
- Grade Group 2: Mostly regular glands with a small component of irregular, fused, and/or cribriform glands.
- Grade Group 3: Mostly irregular, fused, and/or cribriform glands with a small component of regular glands.
- Grade Group 4: Only irregular, fused, and/or cribriform glands, with variations including a minor component of non-glandular tissue.
- Grade Group 5: No glandular formation or presence of necrotic glands with or without irregular, fused, and/or cribriform glands.

The goal of the project was to develop two generative models: the first one aims to simulate semantic segmentation masks of the prostate adenocarcinoma tissue through the implementation of a class-conditioned GAN, while the second one allows for the synthesis of prostate histopathological images from a given ground truth, performing semantic synthesis.

6.7.2. Dataset description and preprocessing

For this project, a dataset provided by the Vancouver Prostate Centre was utilized. This dataset comprises 6 tissue microarrays, each consisting of a series of tissue samples, totaling 333 Whole Slide Images (WSIs). These samples were collected from 231 patients who underwent radical prostatectomy at the Vancouver hospital from June 1997 to June 2011. Each slide was stained with hematoxylin and eosin and digitized using an SCN400 scanner (Leica Microsystems) at a 40x resolution. Image sizes range from 4608x4608 to 5632x5632 pixels. Six pathologists, with an average of 16 years of experience in classifying prostate carcinoma, annotated the WSIs, assigning one of the following five classes to each tissue region: Benign, Gleason Score 3, Gleason Score 4, Gleason Score 5, and Other. Specifically, four pathologists annotated all samples, while the remaining two labeled 191 and 92 images, respectively. These annotations were made between December 2016 and October 2017 [230].

Given the dataset consisted of multiple annotations, a technique was applied to combine the various semantic masks from each expert to create a highly accurate and robust ground truth label. This was achieved by implementing the STAPLE algorithm [231], a weighted voting mechanism that calculates both a probabilistic estimate of the "real" segmentation and a performance measure for each segmentation. This estimate is derived by optimally combining the segmentations, weighing each one

based on estimated performance levels, and incorporating a model for the spatial distribution of segmented structures and spatial homogeneity constraints. Thus, it accounts not only for systematic deviation due to differences in manual annotations but also evaluates the quality of each expert's annotations. The output is an image composed of pixels with numerical values between 0 and 1, representing the probability that each pixel belongs to a specific semantic label. However, in this work, the STAPLE algorithm was adapted for multi-label maps. If, because of STAPLE, some areas were assigned to multiple labels, the majority voting technique was applied. An

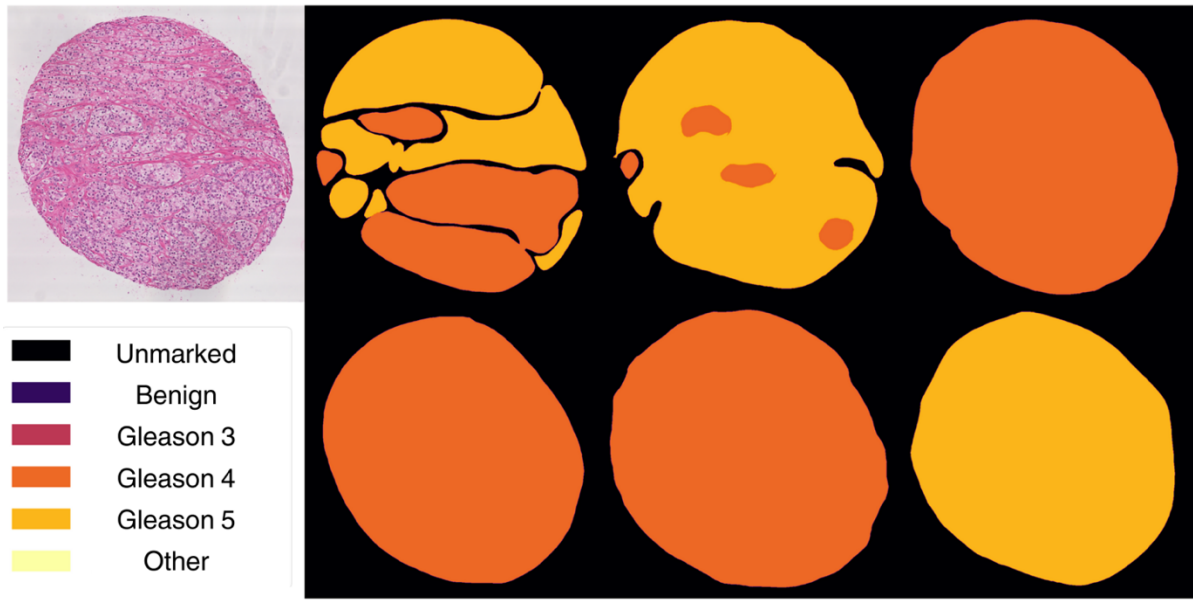


Figure 36 Different readers evaluations of the same WSI.

example of a ground truth label obtained by merging available manual annotations is illustrated in Figure 36.

Using specific algorithms for optimization, normalization, and segmentation of images, optimized and normalized WSIs were first obtained, and from these, the following masks were derived:

Whites: Binary masks identifying parts of the image belonging to the slide, excluding stroma and nuclei. Small artifacts were removed from these masks.

Nuclei: Grayscale masks differentiating nuclei (pixels at 255) and their borders (pixels at 127) from the rest of the image.

These masks were used for two purposes: integrating whites and nuclei within semantic maps and creating instance maps. Firstly, ground truth labels were modified to differentiate image regions belonging to the background from non-marked tissue areas due to uncertain classification. For each WSI, a background mask was derived from the white mask and integrated into the corresponding semantic map. Consequently, the classification of classes was modified as follows: Background, Benign, Gleason 3, Gleason 4, Gleason 5, Unmarked/Other. To better guide the synthesis process in locating nuclei and whites, two new classes were added to enrich the semantic information contained in the segmentation maps, increasing the number of semantic labels to eight: Background, Benign, Gleason 3, Gleason 4, Gleason 5, Nuclei, Whites, Unmarked/Other.

Secondly, so-called instance maps were created, essentially binary masks containing the contours of nuclei and whites. These will be input to the GAN along with the semantic segmentation maps only in

the task of semantic synthesis of histological images, to distinctly differentiate nuclei and whites from the rest of the image.

Since image dimensions were too large to be input to generative models, patch extraction was performed selecting an input size of 512x512. The extraction parameters of patches play a crucial role as they determine the degree of information contained in different patches and the number of elements obtained. Histological features of tissue can be significant at various resolutions, so it's crucial that the set size scale allows for the extraction of patches containing the necessary architectural features for their synthesis. Larger, lower-scale patches show less localized histological features, while smaller, higher-scale patches may not contain enough semantic information.

Overlap between adjacent patches directly regulates the number of extracted patches per image. High overlap leads to a highly repetitive dataset and significant computational time for extraction. Conversely, too low overlap results in an insufficient number of examples. To balance these effects, an overlap value of 50% and a patch size of 512x512 were set. Before applying the sliding window and dividing into patches, data was downsampled by a factor of 2, obtaining images at 20x magnification.

For the synthesis of histopathological segmentation maps, a specific extraction condition was applied. Only patches containing at least 90% of the same class among "Benign," "Gleason 3," "Gleason 4," and "Gleason 5" were selected. For this task, a dataset consisting solely of the training set was used, comprising a total of 8000 grayscale segmentation maps, 2000 for each class. Since there were only 527 images in the "Gleason 5" group, data augmentation was performed, applying transformations such as 90°, 180°, 270° rotation, horizontal flip, and vertical flip.

To simplify the generation task, pixel values associated with each label were converted to utilize the full image dynamic range. Thus, different semantic areas were categorized to maximize the distance between pixels belonging to different labels, thereby maximizing contrast between histological structures.

Unlike the first step, the dataset for semantic image generation was divided into Training and Test sets. The distribution of images and their corresponding semantic and instance maps into these groups was random, ensuring the training set comprised 70% of the data available. An example of a patch with its instance and semantic maps from the created dataset is shown in Figure 37. Specific values were assigned to each pixel of the semantic segmentation masks based on their label.

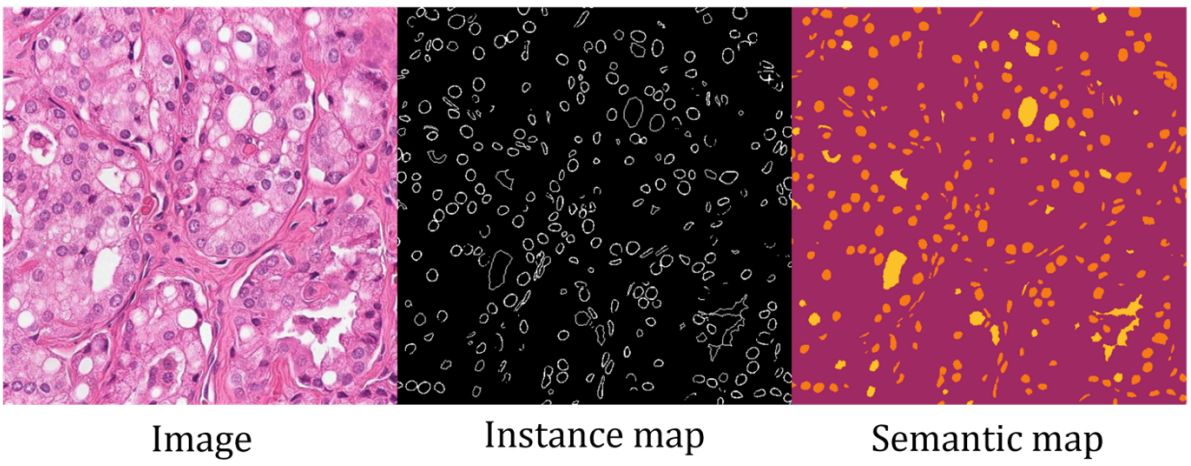


Figure 37 Example of real image and the automatically extracted instance map and segmentation map

6.7.3. Methods

The synthesis of the segmentation maps and the histopathological images were obtained by leveraging two generative models tailored to distinct tasks within the pipeline. For the generation of high-quality individual segmentation maps, StyleGAN2 [77] was employed focusing on the correct representation of object shapes and spatial distribution conditioned on the Gleason class. SPADE was utilized for the synthesis of histopathological images based on the semantic segmentation maps. Comprehensive coverage of both broad and targeted image generation needs within the medical domain was allowed by this bifurcated approach.

Enhanced image synthesis was introduced by StyleGAN2, which builds on the principles of Generative Adversarial Networks (GANs), incorporating several key improvements. A modified architecture that includes adaptive instance normalization (AdaIN) and progressive training techniques was employed to improve convergence and image quality. Images were generated by mapping points in a latent space to detailed, diverse images, underpinned mathematically by a loss function that minimizes the difference between the distributions of generated and real images, ensuring the diversity and reflection of the synthetic images on the target dataset.

Instead of introducing SPADE, its utilization in this context is justified by the need for high-detail and contextually accurate image synthesis, which aligns closely with the input segmentation maps. SPADE's ability to adapt normalization parameters spatially across the image ensures that each synthesized image adheres to the desired medical conditions or anatomical features, providing precise control over the structural and semantic elements of the generated images.

The generation of the semantic map for the first task was evaluated through the FID metric and by subjective evaluation. Morphological parameters describing the nuclei and white classes were also computed, since their shape and spatial distribution is directly linked to the Gleason score. The quality of the generated histopathological images in the second task was evaluated using image quality metrics PSNR, and SSIM.

6.7.4. Results

The StyleGAN2, despite challenges typical of GANs training, successfully simulated 5000 segmentation masks for four pathological grades, allowing for a detailed comparison with real masks in terms of morphology. The analysis revealed a good correspondence in fundamental morphological parameters (area, perimeter, eccentricity, and nucleus count) across classes, with some exceptions. Qualitative results are presented in Figure 38.

Focusing on the nuclei class, the area estimates from generated masks for the benign and Gleason 4 classes align closely with the real masks, showing error percentages of approximately 1.47% and 5.53%, respectively. However, for Gleason 3 and 5 classes, the average area values were higher, with error percentages around 16.2% and 25.4%, respectively. Perimeter values were mostly comparable between real and simulated distributions, except for Gleason 5, where an overestimated mean value resulted in a 13.8% error. Eccentricity estimates were generally underestimated, yet the trend across malignancy grades mirrored the reference, with errors less significant in benign and moderate in higher grades. The model also underestimated the nucleus count, particularly in Gleason 5, indicating a tendency of StyleGAN2 to merge close nuclei into a single entity. Regarding the morphology of the whites, the analysis showed that the average values obtained from simulated samples generally adhere to the reference trends, except for perimeter measurements. Perimeter values for benign and Gleason 3 classes closely matched real data, with minimal error percentages. However, for more

severe grades, discrepancies became significant, especially for Gleason 5, indicating a challenge in accurately simulating the complex morphology of severe pathological conditions.

Qualitatively, SPADE is able to correctly replicate the position of the whites and nuclei of the tissue, returning histological images with realistic appearances. On the other hand, the model has difficulty generating red blood cells and some portions of the stroma. In some images simulated through SPADE, the presence of an artifact, manifesting as a repetitive geometric pattern on the surface of the entire sample, is noticeable.

Analyzing the results in detail, an average PSNR value of approximately 16.1 and 16.9 dB for SPADE is observed. As for the SSIM, average values around 0.245 for SPADE are obtained, reaching higher peaks in images with a high content of white areas.

Dividing the dataset by class, the Benign class achieved significantly superior results in terms of image quality, probably due to the morphology and architecture of the tissue, which includes multiple white areas. Specifically, PSNR values of 17.5 dB, SSIM of 0.390 and 0.397 were obtained. Conversely, for the Gleason 3 and Gleason 4 classes, the obtained values are in line with the overall average values. The Gleason 5 class behaves differently, being associated with worse indices compared to the global averages. In particular, PSNR and SSIM values of 15.95 dB and 0.199 are observed. This difference might also be associated with the complexity and scarcity of Gleason 5 grade samples within the dataset. Figure 39 reports qualitative results on the test set.

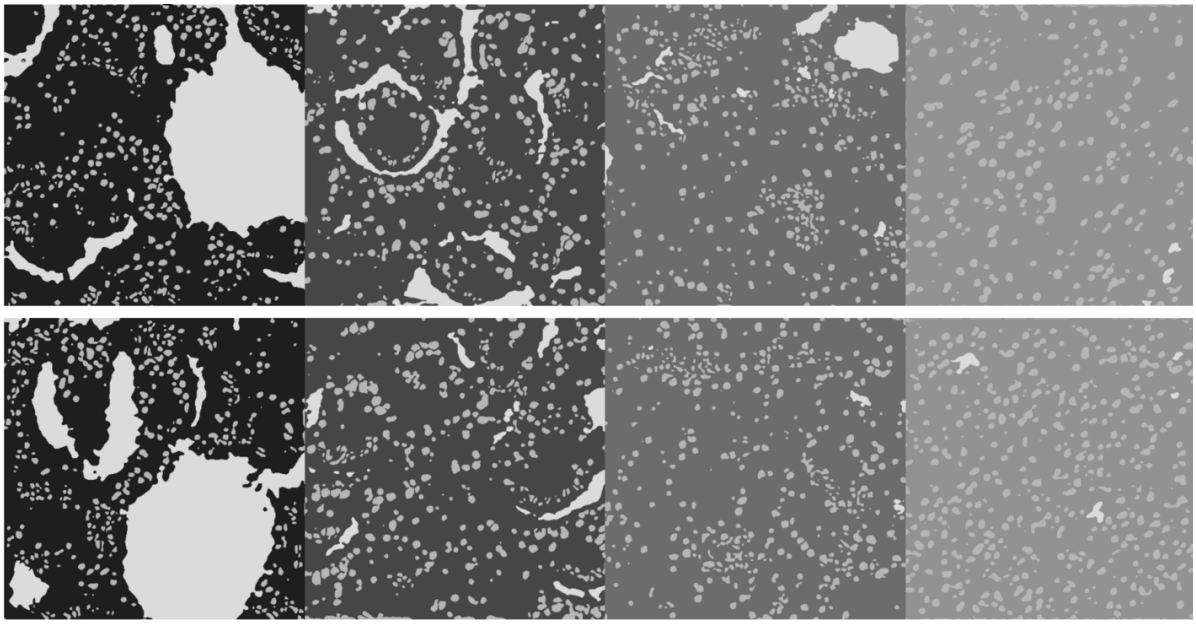


Figure 38 Qualitative result of the simulation of realistic segmentation maps. First row, real images. Second row, generated images.

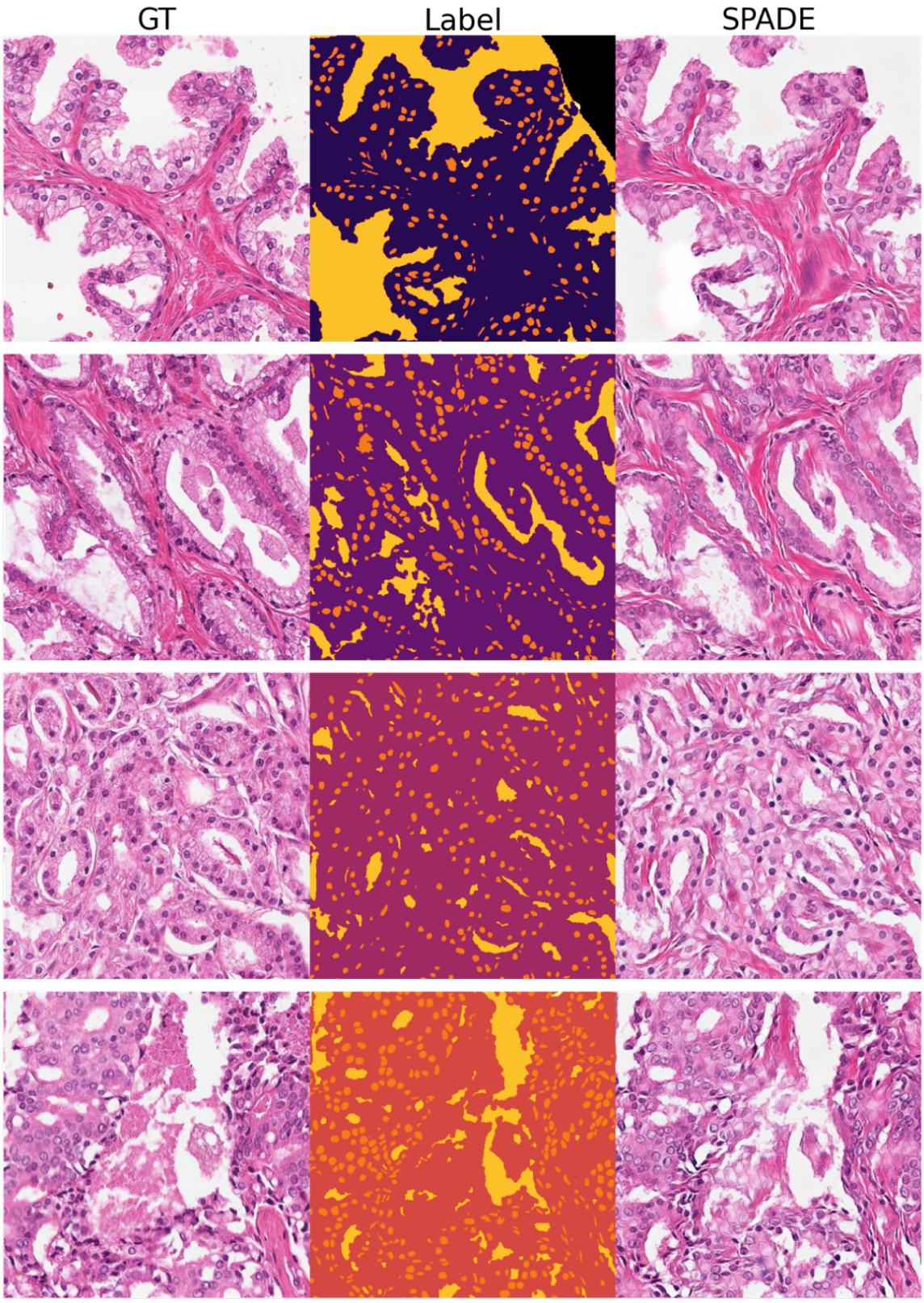


Figure 39 Qualitative results of the histopathological image synthesis using SPADE from segmentation maps.

6.8. Discussion

The integration of the Dual Pyramid Generative Adversarial Network (DP-GAN) within the modular pipeline represents a key advancement for the synthesis of ultrasound images. It enables the of semantic and stylistic uncoupling in a limited data environment. Previous methods like SPADE could correctly represent the texture inside each label, but the synthesized images were clearly identifiable by the sharp transitions between different labels. This limitation has been overcome by exploiting the spatial adaptive normalization of the DP-GAN that preserve the texture representation but can smooth transitions between labels.

The strengths and applications of the current approach are manifold. Primarily, the modular nature of the pipeline allows for flexibility in the generation process, enabling the synthesis of images from diverse inputs including real data, manually defined structures, or AI-generated semantic content. This versatility is particularly helpful in medical imaging, where the need for high-quality, diverse datasets is required to approximate the complex distribution of real data. Moreover, the application of DP-GAN in this pipeline enhances the fidelity and realism of generated images, addressing a critical requirement for effective deep learning model training and validation. The potential of this approach extends beyond data augmentation; it also offers promising opportunities for decision support systems and training tools for medical professionals, thereby contributing to improved diagnostic and therapeutic outcomes.

However, the current approach is not without its limitations. One of the primary challenges lies in the balance between model complexity and computational efficiency. While DP-GAN offers significant advantages in image quality and detail preservation, the computational demands of training and inference can be substantial, especially when dealing with large datasets or aiming for real-time application scenarios. Furthermore, the reliance on high-quality, annotated datasets for initial training remains a significant hurdle, although lessened by the generative capabilities of the model. Possible improvements to this approach could involve the exploration of more efficient model architectures or the incorporation of transfer learning techniques to mitigate the need for extensive annotated datasets. Additionally, advancing the model's ability to handle variability in image quality and acquisition parameters across different ultrasound devices could further enhance its applicability and generalizability.

General insights on generative models in ultrasound imaging suggest a rapidly evolving field with significant potential for impact on clinical practice and research. The progress in this domain highlights the increasing capability of deep learning models not only to generate synthetic data but also to understand and replicate complex biological and anatomical structures. This capability opens new horizons for non-invasive diagnostics, personalized medicine, and the development of advanced therapeutic strategies.

In conclusion, the current approach, centered around the integration of DP-GAN within a modular pipeline for ultrasound image synthesis, presents a promising framework for overcoming some of the challenges in medical imaging. While acknowledging the limitations and potential for further improvements, this work contributes valuable insights into the application of generative models in ultrasound imaging.

7. Conclusion

This thesis explored the transformative potential of deep learning (DL) within medical ultrasound imaging and its extension into broader medical image analysis. The primary outcomes of this research underscore significant potential in the application of DL techniques for discriminative analysis of vascular ultrasound imaging, optic nerve diameter measurement, and musculoskeletal ultrasound imaging. New training and inference techniques, loss functions, and end-to-end pipelines were proposed and discussed in their benefits and limitations.

Expanding on this analysis, a modular pipeline for the controllable synthesis of medical images was proposed. Implementing this innovative framework could potentially improve the quality, diversity, and representativeness of the medical ultrasound datasets, improving the accuracy and reliability of automated systems for the analysis of ultrasound images.

Beyond the enhancement of existing images and datasets, the control over image generation opens new possibilities for utilizing generated images in validating discriminative models and, in creating teaching tools for physician training. A dataset of high quality and realism, generated from a known ground truth, can serve as a benchmark for evaluating discriminative models. This capability to generate images representing various levels of pathology or rare diagnostic conditions could prove invaluable in medical education. Physicians, especially those in the early stages of their careers, could significantly benefit from exposure to a broader spectrum of cases, including those they might rarely encounter in clinical practice.

A critical challenge that future work must address is the definition of image quality metrics and benchmarks for generated images. Ensuring the safety and correctness of generated images is crucial, particularly when these images are intended for use in clinical decision-making or medical education. The development of standardized, rigorous metrics for assessing the quality, realism, and clinical relevance of synthetic images will be the key step to make generative AI applicable in medical imaging. These metrics must account not only for visual fidelity but also for the anatomical and pathological accuracy of the images, ensuring that they are a reliable basis for training, evaluation, and potentially even diagnostic interpretation.

In conclusion, the journey of this Ph.D. research has been towards the enhancement of diagnostic precision of automated methods for medical image analysis. It involved a wide variety of methods to achieve this goal, from traditional rule-based models and methods mimicking physician evaluation, to the application of deep learning models both as discriminative and generative tool. Eventually, this thesis tries to chart a path forward for the application of deep learning in medical imaging to ensure that the benefits of AI-driven technologies are realized in the service of patient care and medical education.

List of publications

Papers

Marzola, Francesco; Lochner, Piergiorgio; Naldi, Andrea; Lemor, Robert; Stögbauer, Jakob; Meiburger, Kristen M; Development of a Deep Learning–Based System for Optic Nerve Characterization in Transorbital Ultrasound Images on a Multicenter Data Set; ULTRASOUND IN MEDICINE AND BIOLOGY; ISSN 0301-5629. 49:9(2023), pp. 2060-2071. [10.1016/j.ultrasmedbio.2023.05.011]

Meiburger, K. M.; **Marzola, Francesco**; Zahnd, G.; Faita, F.; Loizou, C. P.; Laine, N.; Carvalho, C.; Steinman, D. A.; Gibello, L.; Bruno, R. M.; Clarenbach, R.; Francesconi, M.; Nicolaides, A. N.; Liebgott, H.; Campilho, A.; Ghotbi, R.; Kyriacou, E.; Navab, N.; Griffin, M.; Panayiotou, A. G.; Gherardini, R.; Varetto, G.; Bianchini, E.; Pattichis, C. S.; Ghiadoni, L.; Rouco, J.; Orkisz, M.; Molinari, F.; Carotid Ultrasound Boundary Study (CUBS): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal B-mode ultrasound scans; COMPUTERS IN BIOLOGY AND MEDICINE; ISSN 0010-4825. 144:(2022), p. 105333. [10.1016/j.compbiomed.2022.105333]

Marzola, Francesco; van Alfen, Nens; Doorduin, Jonne; Meiburger, Kristen M.; Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment; COMPUTERS IN BIOLOGY AND MEDICINE; ISSN 0010-4825. 135:(2021); p. 104623. [10.1016/j.compbiomed.2021.104623]

Meiburger, K. M.; Zahnd, G.; Faita, F.; Loizou, C. P.; Carvalho, C.; Steinman, D. A.; Gibello, L.; Bruno, R. M.; **Marzola, Francesco**; Clarenbach, R.; Francesconi, M.; Nicolaides, A. N.; Campilho, A.; Ghotbi, R.; Kyriacou, E.; Navab, N.; Griffin, M.; Panayiotou, A. G.; Gherardini, R.; Varetto, G.; Bianchini, E.; Pattichis, C. S.; Ghiadoni, L.; Rouco, J.; Molinari, F.; Carotid Ultrasound Boundary Study (CUBS): An Open Multicenter Analysis of Computerized Intima–Media Thickness Measurement Systems and Their Clinical Impact; ULTRASOUND IN MEDICINE AND BIOLOGY; ISSN 0301-5629.; 47:8(2021), pp. 2442-2455. [10.1016/j.ultrasmedbio.2021.03.022]

Proceedings

Marzola, Francesco; Meiburger, Kristen M.; Salvi, Massimo. Innovative temporal loss function for segmentation of fine structures in ultrasound images; (2023); (2023 IEEE International Ultrasonics Symposium (IUS) Montreal, QC, Canada 03-08 September 2023) [10.1109/IUS51837.2023.10308305].

Marzola, Francesco; Meiburger, Kristen; Molinari, Filippo; Salvi, Massimo. Can multiple segmentation methods enhance deep learning networks generalization? A novel hybrid learning paradigm; 12465:(2023), p. 39. (SPIE MEDICAL IMAGING 2023 tenutosi a San Diego (USA) nel 19-24 Feb 2023) [10.1117/12.2653394].

Marzola, Francesco; Meiburger, Kristen M.; Molinari, Filippo; Salvi, Massimo. Exploring the Impact of Learning Paradigms on Network Generalization: A Multi-Center IMT Study; (2023), pp. 1-5. (24th International Conference on Digital Signal Processing (DSP) Rhodes (Rodos), Greece 11-13 June 2023) [10.1109/DSP58604.2023.10167892].

Marzola, Francesco; Meiburger, Kristen M.; Naldi, Andrea; Lochner, Piergiorgio; Automatic segmentation of the optic nerve in transorbital ultrasound images using a deep learning approach; (2021), pp. 1-4. (2021 IEEE International Ultrasonics Symposium (IUS) Virtual 11-16 Sept. 2021) [10.1109/IUS52206.2021.9593827].

Marzola, Francesco; Alfen, N. V.; Salvi, M.; De Santi, B.; Doorduin, J.; Meiburger, K. M.; Automatic segmentation of ultrasound images of gastrocnemius medialis with different echogenicity levels using convolutional neural networks; (2020), pp. 2113-2116. (42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, EMBC 2020 a Montreal, QC, Canada 20-24 July 2020) [10.1109/EMBC44109.2020.9176343].

De Santi, B.; Salvi, M.; Giannini, V.; Meiburger, K. M.; **Marzola, F**.; Russo, F.; Bosco, M.; Molinari, F.; Comparison of Histogram-based Textural Features between Cancerous and Normal Prostatic Tissue in Multiparametric

Magnetic Resonance Images; (2020), pp. 1671-1674. (42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, EMBC 2020) [10.1109/EMBC44109.2020.9176307].

Bibliography

[1] C. M. Rumack, *Diagnostic ultrasound*, 4th ed. Philadelphia, PA: Elsevier/Mosby, 2011.

[2] *Diagnostic Ultrasound Imaging: Inside Out*. Elsevier, 2014. doi: 10.1016/C2011-0-07261-7.

[3] J. M. Sanches, A. Laine, and J. S. Suri, *Ultrasound imaging: advances and applications*. New York, NY: Springer, 2012.

[4] Y. Sharifi, M. A. Bakhshali, T. Dehghani, M. DanaiAshgzari, M. Sargolzaei, and S. Eslami, "Deep learning on ultrasound images of thyroid nodules," *Biocybern. Biomed. Eng.*, vol. 41, no. 2, pp. 636–655, Apr. 2021, doi: 10.1016/j.bbe.2021.02.008.

[5] G.-G. Wu *et al.*, "Artificial intelligence in breast ultrasound," *World J. Radiol.*, vol. 11, no. 2, pp. 19–26, Feb. 2019, doi: 10.4329/wjr.v11.i2.19.

[6] M. Lupsor-Platon, T. Serban, A. I. Silion, G. R. Tirpe, A. Tirpe, and M. Florea, "Performance of Ultrasound Techniques and the Potential of Artificial Intelligence in the Evaluation of Hepatocellular Carcinoma and Non-Alcoholic Fatty Liver Disease," *Cancers*, vol. 13, no. 4, p. 790, Feb. 2021, doi: 10.3390/cancers13040790.

[7] R. Horgan, L. Nehme, and A. Abuhamad, "Artificial intelligence in obstetric ultrasound: A scoping review," *Prenat. Diagn.*, vol. 43, no. 9, pp. 1176–1219, Aug. 2023, doi: 10.1002/pd.6411.

[8] K. A. Lara Hernandez, T. Rienmüller, D. Baumgartner, and C. Baumgartner, "Deep learning in spatiotemporal cardiac imaging: A review of methodologies and clinical usability," *Comput. Biol. Med.*, vol. 130, p. 104200, Mar. 2021, doi: 10.1016/j.compbiomed.2020.104200.

[9] Y. Shin, J. Yang, Y. H. Lee, and S. Kim, "Artificial intelligence in musculoskeletal ultrasound imaging," *Ultrasonography*, vol. 40, no. 1, pp. 30–44, Jan. 2021, doi: 10.14366/usg.20080.

[10] M. Komatsu *et al.*, "Towards Clinical Application of Artificial Intelligence in Ultrasound Imaging," *Biomedicines*, vol. 9, no. 7, p. 720, Jun. 2021, doi: 10.3390/biomedicines9070720.

[11] B. Luijten, N. Chennakeshava, Y. C. Eldar, M. Mischi, and R. J. G. Van Sloun, "Ultrasound Signal Processing: From Models to Deep Learning," *Ultrasound Med. Biol.*, vol. 49, no. 3, pp. 677–698, Mar. 2023, doi: 10.1016/j.ultrasmedbio.2022.11.003.

[12] K. Suzuki, "Overview of deep learning in medical imaging," *Radiol. Phys. Technol.*, vol. 10, no. 3, pp. 257–273, Sep. 2017, doi: 10.1007/s12194-017-0406-5.

[13] J.-G. Lee *et al.*, "Deep Learning in Medical Imaging: General Overview," *Korean J. Radiol.*, vol. 18, no. 4, p. 570, 2017, doi: 10.3348/kjr.2017.18.4.570.

[14] M. Kim *et al.*, "Deep Learning in Medical Imaging," *Neurospine*, vol. 16, no. 4, pp. 657–668, Dec. 2019, doi: 10.14245/ns.1938396.198.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, doi: 10.48550/ARXIV.1512.03385.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2016, doi: 10.48550/ARXIV.1608.06993.

[18] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," 2017, doi: 10.48550/ARXIV.1709.01507.

[19] A. Vaswani *et al.*, "Attention Is All You Need," 2017, doi: 10.48550/ARXIV.1706.03762.

[20] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, doi: 10.48550/ARXIV.2010.11929.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[22] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 2015, doi: 10.48550/ARXIV.1502.03167.

[23] R. Müller, S. Kornblith, and G. Hinton, "When Does Label Smoothing Help?," 2019, doi: 10.48550/ARXIV.1906.02629.

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," 2017, doi: 10.48550/ARXIV.1710.09412.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," 2020, doi: 10.48550/ARXIV.2002.05709.

[26] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, “Self-Supervised Representation Learning: Introduction, advances, and challenges,” *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, May 2022, doi: 10.1109/MSP.2021.3134634.
[27] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 2014, doi: 10.48550/ARXIV.1412.6980.
[28] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, “Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network,” *Appl. Intell.*, vol. 51, no. 2, pp. 854–864, Feb. 2021, doi: 10.1007/s10489-020-01829-7.
[29] S. K. Thakur, D. P. Singh, and J. Choudhary, “Lung cancer identification: a review on detection and classification,” *Cancer Metastasis Rev.*, vol. 39, no. 3, pp. 989–998, Sep. 2020, doi: 10.1007/s10555-020-09901-x.
[30] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, “Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review,” *Diagnostics*, vol. 11, no. 8, p. 1390, Jul. 2021, doi: 10.3390/diagnostics11081390.
[31] S. Benyahia, B. Meftah, and O. Lézoray, “Multi-features extraction based on deep learning for skin lesion classification,” *Tissue Cell*, vol. 74, p. 101701, Feb. 2022, doi: 10.1016/j.tice.2021.101701.
[32] F. Veronese *et al.*, “Teledermoscopy in the Diagnosis of Melanocytic and Non-Melanocytic Skin Lesions: NurugoTM Derma Smartphone Microscope as a Possible New Tool in Daily Clinical Practice,” *Diagnostics*, vol. 12, no. 6, p. 1371, Jun. 2022, doi: 10.3390/diagnostics12061371.
[33] N. Tsiknakis *et al.*, “Deep learning for diabetic retinopathy detection and classification based on fundus images: A review,” *Comput. Biol. Med.*, vol. 135, p. 104599, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104599.
[34] T. Li *et al.*, “Applications of deep learning in fundus images: A review,” *Med. Image Anal.*, vol. 69, p. 101971, Apr. 2021, doi: 10.1016/j.media.2021.101971.
[35] H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep Learning Approach for Early Detection of Alzheimer’s Disease,” *Cogn. Comput.*, vol. 14, no. 5, pp. 1711–1727, Sep. 2022, doi: 10.1007/s12559-021-09946-2.
[36] G. Liang, H. Hong, W. Xie, and L. Zheng, “Combining Convolutional Neural Network With Recursive Neural Network for Blood Cell Image Classification,” *IEEE Access*, vol. 6, pp. 36188–36197, 2018, doi: 10.1109/ACCESS.2018.2846685.
[37] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, “The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis,” *Comput. Biol. Med.*, vol. 128, p. 104129, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104129.
[38] Z. Akkus *et al.*, “A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow,” *J. Am. Coll. Radiol.*, vol. 16, no. 9, pp. 1318–1328, Sep. 2019, doi: 10.1016/j.jacr.2019.06.004.
[39] J. Ma, F. Wu, T. Jiang, J. Zhu, and D. Kong, “Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images,” *Med. Phys.*, vol. 44, no. 5, pp. 1678–1691, May 2017, doi: 10.1002/mp.12134.
[40] S. Han *et al.*, “A deep learning framework for supporting the classification of breast lesions in ultrasound images,” *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7714–7728, Sep. 2017, doi: 10.1088/1361-6560/aa82ec.
[41] X. Liu, J. Song, S. Wang, J. Zhao, and Y. Chen, “Learning to Diagnose Cirrhosis with Liver Capsule Guided Ultrasound Image Classification,” *Sensors*, vol. 17, no. 12, p. 149, Jan. 2017, doi: 10.3390/s17010149.
[42] Z. Yu *et al.*, “A Deep Convolutional Neural Network-Based Framework for Automatic Fetal Facial Standard Plane Recognition,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 3, pp. 874–885, May 2018, doi: 10.1109/JBHI.2017.2705031.
[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-Based Convolutional Networks for Accurate Object Detection and Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.
[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2015, doi: 10.48550/ARXIV.1506.02640.
[45] K. Liu, “STBi-YOLO: A Real-Time Object Detection Method for Lung Nodule Recognition,” *IEEE Access*, vol. 10, pp. 75385–75394, 2022, doi: 10.1109/ACCESS.2022.3192034.
[46] N. M. Dipu, S. A. Shohan, and K. M. A. Salam, “Deep Learning Based Brain Tumor Detection and Classification,” in *2021 International Conference on Intelligent Technologies (CONIT)*, Hubli, India: IEEE, Jun. 2021, pp. 1–6. doi: 10.1109/CONIT51480.2021.9498384.
[47] S. Zhang, S. Xu, L. Tan, H. Wang, and J. Meng, “Stroke Lesion Detection and Analysis in MRI Images Based on Deep Learning,” *J. Healthc. Eng.*, vol. 2021, pp. 1–9, Apr. 2021, doi: 10.1155/2021/5524769.

[48] S. Ito *et al.*, "Automated Detection of Spinal Schwannomas Utilizing Deep Learning Based on Object Detection From Magnetic Resonance Imaging," *Spine*, vol. 46, no. 2, pp. 95–100, Jan. 2021, doi: 10.1097/BRS.0000000000003749.
[49] C. F. Baumgartner *et al.*, "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound," *IEEE Trans. Med. Imaging*, vol. 36, no. 11, pp. 2204–2215, Nov. 2017, doi: 10.1109/TMI.2017.2712367.
[50] M. H. Yap *et al.*, "Breast ultrasound region of interest detection and lesion localisation," *Artif. Intell. Med.*, vol. 107, p. 101880, Jul. 2020, doi: 10.1016/j.artmed.2020.101880.
[51] F. Abdolali, J. Kapur, J. L. Jaremko, M. Noga, A. R. Hareendranathan, and K. Punithakumar, "Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks," *Comput. Biol. Med.*, vol. 122, p. 103871, Jul. 2020, doi: 10.1016/j.compbiomed.2020.103871.
[52] X. Wu *et al.*, "CacheTrack-YOLO: Real-Time Detection and Tracking for Thyroid Nodules and Surrounding Tissues in Ultrasound Videos," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 10, pp. 3812–3823, Oct. 2021, doi: 10.1109/JBHI.2021.3084962.
[53] A. F. Al-Battal *et al.*, "A CNN Segmentation-Based Approach to Object Detection and Tracking in Ultrasound Scans with Application to the Vagus Nerve Detection," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Mexico: IEEE, Nov. 2021, pp. 3322–3327. doi: 10.1109/EMBC46164.2021.9630522.
[54] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," 2014, doi: 10.48550/ARXIV.1411.4038.
[55] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, doi: 10.48550/ARXIV.1505.04597.
[56] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," 2018, doi: 10.48550/ARXIV.1807.10165.
[57] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," 2017, doi: 10.48550/ARXIV.1706.05587.
[58] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," 2021, doi: 10.48550/ARXIV.2103.10504.
[59] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022, doi: 10.1049/ipr2.12419.
[60] X. Liu, L. Song, S. Liu, and Y. Zhang, "A Review of Deep-Learning-Based Medical Image Segmentation Methods," *Sustainability*, vol. 13, no. 3, p. 1224, Jan. 2021, doi: 10.3390/su13031224.
[61] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.
[62] S. Liu *et al.*, "Deep Learning in Medical Ultrasound Analysis: A Review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019, doi: 10.1016/j.eng.2018.11.020.
[63] Y. Wang, X. Ge, H. Ma, S. Qi, G. Zhang, and Y. Yao, "Deep Learning in Medical Ultrasound Image Analysis: A Review," *IEEE Access*, vol. 9, pp. 54310–54324, 2021, doi: 10.1109/ACCESS.2021.3071301.
[64] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," 2016, doi: 10.48550/ARXIV.1610.02391.
[66] J. Heo, S. Joo, and T. Moon, "Fooling Neural Network Interpretations via Adversarial Model Manipulation," 2019, doi: 10.48550/ARXIV.1902.02041.
[67] S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)," *Comput. Biol. Med.*, vol. 165, p. 107441, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107441.
[68] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.
[69] H. Shao, A. Kumar, and P. T. Fletcher, "The Riemannian Geometry of Deep Generative Models," 2017, doi: 10.48550/ARXIV.1711.08014.
[70] S. Arora and Y. Zhang, "Do GANs actually learn the distribution? An empirical study," 2017, doi: 10.48550/ARXIV.1706.08224.
[71] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," 2014, doi: 10.48550/ARXIV.1406.2661.
[72] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," 2016, doi: 10.48550/ARXIV.1606.03498.

[73] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, doi: 10.48550/ARXIV.1701.07875.
[74] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, “Improving GANs Using Optimal Transport,” 2018, doi: 10.48550/ARXIV.1803.05573.
[75] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” 2015, doi: 10.48550/ARXIV.1511.06434.
[76] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” 2017, doi: 10.48550/ARXIV.1710.10196.
[77] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” 2019, doi: 10.48550/ARXIV.1912.04958.
[78] M. Kang *et al.*, “Scaling up GANs for Text-to-Image Synthesis,” 2023, doi: 10.48550/ARXIV.2303.05511.
[79] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” 2016, doi: 10.48550/ARXIV.1611.07004.
[80] B. Dai, S. Fidler, R. Urtasun, and D. Lin, “Towards Diverse and Natural Image Descriptions via a Conditional GAN,” 2017, doi: 10.48550/ARXIV.1703.06029.
[81] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” 2017, doi: 10.48550/ARXIV.1703.10593.
[82] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using GAN for improved liver lesion classification,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC: IEEE, Apr. 2018, pp. 289–293. doi: 10.1109/ISBI.2018.8363576.
[83] Y. Sun, P. Yuan, and Y. Sun, “MM-GAN: 3D MRI Data Augmentation for Medical Image Segmentation via Generative Adversarial Networks,” in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, Nanjing, China: IEEE, Aug. 2020, pp. 227–234. doi: 10.1109/ICBK50248.2020.00041.
[84] T. Pang, J. H. D. Wong, W. L. Ng, and C. S. Chan, “Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification,” *Comput. Methods Programs Biomed.*, vol. 203, p. 106018, May 2021, doi: 10.1016/j.cmpb.2021.106018.
[85] Y. Ma *et al.*, “Structure and Illumination Constrained GAN for Medical Image Enhancement,” *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3955–3967, Dec. 2021, doi: 10.1109/TMI.2021.3101937.
[86] M. Salvi *et al.*, “DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks,” *Comput. Methods Programs Biomed.*, vol. 225, p. 107040, Oct. 2022, doi: 10.1016/j.cmpb.2022.107040.
[87] M. Salvi, F. Branciforti, F. Molinari, and K. M. Meiburger, “Generative models for color normalization in digital pathology and dermatology: Advancing the learning paradigm,” *Expert Syst. Appl.*, vol. 245, p. 123105, Jul. 2024, doi: 10.1016/j.eswa.2023.123105.
[88] P. Welander, S. Karlsson, and A. Eklund, “Generative Adversarial Networks for Image-to-Image Translation on Multi-Contrast MR Images - A Comparison of CycleGAN and UNIT,” 2018, doi: 10.48550/ARXIV.1806.07777.
[89] A. Abu-Srhan, I. Almallahi, M. A. M. Abushariah, W. Mahafza, and O. S. Al-Kadi, “Paired-unpaired Unsupervised Attention Guided GAN with transfer learning for bidirectional brain MR-CT synthesis,” *Comput. Biol. Med.*, vol. 136, p. 104763, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104763.
[90] D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” *Found. Trends® Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.
[91] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” 2013, doi: 10.48550/ARXIV.1312.6114.
[92] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, “Normalizing Flows: An Introduction and Review of Current Methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3964–3979, Nov. 2021, doi: 10.1109/TPAMI.2020.2992934.
[93] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” 2020, doi: 10.48550/ARXIV.2006.11239.
[94] G. A. Roth *et al.*, “Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019,” *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020, doi: 10.1016/j.jacc.2020.11.010.
[95] G. A. Mensah, G. A. Roth, and V. Fuster, “The Global Burden of Cardiovascular Diseases and Risk Factors,” *J. Am. Coll. Cardiol.*, vol. 74, no. 20, pp. 2529–2532, Nov. 2019, doi: 10.1016/j.jacc.2019.10.009.
[96] P. Libby, “The changing landscape of atherosclerosis,” *Nature*, vol. 592, no. 7855, pp. 524–533, Apr. 2021, doi: 10.1038/s41586-021-03392-8.
[97] P. Langhorne, J. Bernhardt, and G. Kwakkel, “Stroke rehabilitation,” *The Lancet*, vol. 377, no. 9778, pp. 1693–1702, May 2011, doi: 10.1016/S0140-6736(11)60325-5.

[98] A. Manbachi, Y. Hoi, B. A. Wasserman, E. G. Lakatta, and D. A. Steinman, "On the shape of the common carotid artery with implications for blood velocity profiles," *Physiol. Meas.*, vol. 32, no. 12, pp. 1885–1897, Dec. 2011, doi: 10.1088/0967-3334/32/12/001.
[99] E. G. Grant *et al.*, "Carotid Artery Stenosis: Gray-Scale and Doppler US Diagnosis—Society of Radiologists in Ultrasound Consensus Conference," *Radiology*, vol. 229, no. 2, pp. 340–346, Nov. 2003, doi: 10.1148/radiol.2292030516.
[100] P.-J. Touboul *et al.*, "Mannheim Carotid Intima-Media Thickness and Plaque Consensus (2004–2006–2011)," *Cerebrovasc. Dis.*, vol. 34, no. 4, pp. 290–296, 2012, doi: 10.1159/000343145.
[101] L. Engelen, I. Ferreira, C. D. Stehouwer, P. Boutouyrie, S. Laurent, and on behalf of the Reference Values for Arterial Measurements Collaboration, "Reference intervals for common carotid intima-media thickness measured with echotracking: relation with risk factors," *Eur. Heart J.*, vol. 34, no. 30, pp. 2368–2380, Aug. 2013, doi: 10.1093/eurheartj/ehs380.
[102] M. Plichart *et al.*, "Carotid intima-media thickness in plaque-free site, carotid plaques and coronary heart disease risk prediction in older adults. The Three-City Study," *Atherosclerosis*, vol. 219, no. 2, pp. 917–924, Dec. 2011, doi: 10.1016/j.atherosclerosis.2011.09.024.
[103] M. L. Bots *et al.*, "Common Carotid Intima-Media Thickness Measurements Do Not Improve Cardiovascular Risk Prediction in Individuals With Elevated Blood Pressure: The USE-IMT Collaboration," *Hypertension*, vol. 63, no. 6, pp. 1173–1181, Jun. 2014, doi: 10.1161/HYPERTENSIONAHA.113.02683.
[104] M. W. Lorenz *et al.*, "Correction: Predictive value for cardiovascular events of common carotid intima media thickness and its rate of change in individuals at high cardiovascular risk - Results from the PROG-IMT collaboration," *PLOS ONE*, vol. 13, no. 9, p. e0204633, Sep. 2018, doi: 10.1371/journal.pone.0204633.
[105] K. M. Meiburger *et al.*, "Carotid Ultrasound Boundary Study (CUBS): An Open Multicenter Analysis of Computerized Intima–Media Thickness Measurement Systems and Their Clinical Impact," *Ultrasound Med. Biol.*, vol. 47, no. 8, pp. 2442–2455, Aug. 2021, doi: 10.1016/j.ultrasmedbio.2021.03.022.
[106] F. Molinari *et al.*, "Constrained snake vs. conventional snake for carotid ultrasound automated IMT measurements on multi-center data sets," *Ultrasonics*, vol. 52, no. 7, pp. 949–961, Sep. 2012, doi: 10.1016/j.ultras.2012.03.005.
[107] C. P. Loizou, C. S. Pattichis, M. Pantziaris, T. Tyllis, and A. Nicolaides, "Snakes based segmentation of the common carotid artery intima media," *Med. Biol. Eng. Comput.*, vol. 45, no. 1, pp. 35–49, Jan. 2007, doi: 10.1007/s11517-006-0140-3.
[108] Y. Nagaraj, P. Madipalli, J. Rajan, P. K. Kumar, and A. V. Narasimhadhan, "Segmentation of intima media complex from carotid ultrasound images using wind driven optimization technique," *Biomed. Signal Process. Control*, vol. 40, pp. 462–472, Feb. 2018, doi: 10.1016/j.bspc.2017.08.009.
[109] R. Rocha, J. Silva, and A. Campilho, "Automatic segmentation of carotid B-mode images using fuzzy classification," *Med. Biol. Eng. Comput.*, vol. 50, no. 5, pp. 533–545, May 2012, doi: 10.1007/s11517-012-0883-y.
[110] G. Zahnd *et al.*, "A Fully-Automatic Method to Segment the Carotid Artery Layers in Ultrasound Imaging: Application to Quantify the Compression-Decompression Pattern of the Intima-Media Complex During the Cardiac Cycle," *Ultrasound Med. Biol.*, vol. 43, no. 1, pp. 239–257, Jan. 2017, doi: 10.1016/j.ultrasmedbio.2016.08.016.
[111] Y. Nagaraj, A. Hema Sai Teja, and A. V. Narasimhadhan, "Automatic Segmentation of Intima Media Complex in Carotid Ultrasound Images Using Support Vector Machine," *Arab. J. Sci. Eng.*, vol. 44, no. 4, pp. 3489–3496, Apr. 2019, doi: 10.1007/s13369-018-3549-8.
[112] N. Y., A. C.S., H. S. Teja A., and A. V. Narasimhadhan, "Carotid wall segmentation in longitudinal ultrasound images using structured random forest," *Comput. Electr. Eng.*, vol. 69, pp. 753–767, Jul. 2018, doi: 10.1016/j.compeleceng.2018.02.010.
[113] M. Hassan, A. Chaudhry, A. Khan, and M. A. Iftikhar, "Robust information gain based fuzzy c-means clustering and classification of carotid artery ultrasound images," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 593–609, Feb. 2014, doi: 10.1016/j.cmpb.2013.10.012.
[114] P. K. Jain, N. Sharma, M. K. Kalra, A. Johri, L. Saba, and J. S. Suri, "Far wall plaque segmentation and area measurement in common and internal carotid artery ultrasound using U-series architectures: An unseen Artificial Intelligence paradigm for stroke risk assessment," *Comput. Biol. Med.*, vol. 149, p. 106017, Oct. 2022, doi: 10.1016/j.compbiomed.2022.106017.
[115] P. K. Jain *et al.*, "Automated deep learning-based paradigm for high-risk plaque detection in B-mode common carotid ultrasound scans: an asymptomatic Japanese cohort study," *Int. Angiol.*, vol. 41, no. 1, Feb. 2022, doi: 10.23736/S0392-9590.21.04771-4.

[116] N. Lainé, H. Liebgott, G. Zahnd, and M. Orkisz, “Carotid Artery Wall Segmentation in Ultrasound Image Sequences Using a Deep Convolutional Neural Network,” in *Computer Vision and Graphics*, vol. 598, L. J. Chmielewski and A. Orłowski, Eds., in Lecture Notes in Networks and Systems, vol. 598. , Cham: Springer Nature Switzerland, 2023, pp. 73–84. doi: 10.1007/978-3-031-22025-8_6.
[117] G.-Q. Zhou *et al.*, “BSMNet: Boundary-salience multi-branch network for intima-media identification in carotid ultrasound images,” *Comput. Biol. Med.*, vol. 162, p. 107092, Aug. 2023, doi: 10.1016/j.compbiomed.2023.107092.
[118] Y. Yuan, C. Li, L. Xu, S. Zhu, Y. Hua, and J. Zhang, “CSM-Net: Automatic joint segmentation of intima-media complex and lumen in carotid artery ultrasound images,” *Comput. Biol. Med.*, vol. 150, p. 106119, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106119.
[119] H. Hassen Mohammed *et al.*, “Ultrasound Intima-Media Complex (IMC) Segmentation Using Deep Learning Models,” *Appl. Sci.*, vol. 13, no. 8, p. 4821, Apr. 2023, doi: 10.3390/app13084821.
[120] K. M. Meiburger, U. R. Acharya, and F. Molinari, “Automated localization and segmentation techniques for B-mode ultrasound images: A review,” *Comput. Biol. Med.*, vol. 92, pp. 210–235, Jan. 2018, doi: 10.1016/j.compbiomed.2017.11.018.
[121] R.-M. Menchón-Lara and J.-L. Sancho-Gómez, “Fully automatic segmentation of ultrasound common carotid artery images based on machine learning,” *Neurocomputing*, vol. 151, pp. 161–167, Mar. 2015, doi: 10.1016/j.neucom.2014.09.066.
[122] P. K. Kumar, T. Araki, J. Rajan, J. R. Laird, A. Nicolaides, and J. S. Suri, “State-of-the-art review on automated lumen and adventitial border delineation and its measurements in carotid ultrasound,” *Comput. Methods Programs Biomed.*, vol. 163, pp. 155–168, Sep. 2018, doi: 10.1016/j.cmpb.2018.05.015.
[123] Q. Huang, H. Tian, L. Jia, Z. Li, and Z. Zhou, “A review of deep learning segmentation methods for carotid artery ultrasound images,” *Neurocomputing*, vol. 545, p. 126298, Aug. 2023, doi: 10.1016/j.neucom.2023.126298.
[124] K. M. Meiburger *et al.*, “Carotid Ultrasound Boundary Study (CUBS): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal B-mode ultrasound scans,” *Comput. Biol. Med.*, vol. 144, p. 105333, May 2022, doi: 10.1016/j.compbiomed.2022.105333.
[125] L. Aguilar, J. Wong, D. A. Steinman, and R. S. C. Cobbold, “FAMUS II: A Fast and Mechanistic Ultrasound Simulator Using an Impulse Response Approach,” *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 64, no. 2, pp. 362–373, Feb. 2017, doi: 10.1109/TUFFC.2016.2632706.
[126] L. Aguilar, R. C. Cobbold, and D. Steinman, “Fast and mechanistic ultrasound simulation using a point source/receiver approach,” *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 60, no. 11, pp. 2335–2346, Nov. 2013, doi: 10.1109/TUFFC.2013.6644737.
[127] F. Molinari *et al.*, “Ultrasound IMT measurement on a multi-ethnic and multi-institutional database: Our review and experience using four fully automated and one semi-automated methods,” *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 946–960, Dec. 2012, doi: 10.1016/j.cmpb.2012.05.008.
[128] L. Saba *et al.*, “Evaluation of Carotid Wall Thickness by using Computed Tomography and Semiautomated Ultrasonographic Software,” *J. Vasc. Ultrasound*, vol. 35, no. 3, pp. 136–142, Sep. 2011, doi: 10.1177/154431671103500302.
[129] F. Faita, V. Gemignani, E. Bianchini, C. Giannarelli, L. Ghiadoni, and M. Demi, “Real-time Measurement System for Evaluation of the Carotid Intima-Media Thickness With a Robust Edge Operator,” *J. Ultrasound Med.*, vol. 27, no. 9, pp. 1353–1361, Sep. 2008, doi: 10.7863/jum.2008.27.9.1353.
[130] F. Molinari *et al.*, “Fully Automated Dual-Snake Formulation for Carotid Intima-Media Thickness Measurement: A New Approach,” *J. Ultrasound Med.*, vol. 31, no. 7, pp. 1123–1136, Jul. 2012, doi: 10.7863/jum.2012.31.7.1123.
[131] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” 2015, doi: 10.48550/ARXIV.1506.01497.
[132] H. Jiang and E. Learned-Miller, “Face Detection with the Faster R-CNN,” 2016, doi: 10.48550/ARXIV.1606.03473.
[133] MONAI Consortium, “MONAI: Medical Open Network for AI.” Zenodo, Dec. 19, 2022. doi: 10.5281/ZENODO.7459814.
[134] P. Iakubovskii, “Segmentation Models Pytorch,” *GitHub repository*. GitHub, 2019. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
[135] K. Chen *et al.*, “MMDetection: Open MMLab Detection Toolbox and Benchmark,” 2019, doi: 10.48550/ARXIV.1906.07155.

[136] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.
[137] F. Marzola, K. M. Meiburger, F. Molinari, and M. Salvi, "Exploring the Impact of Learning Paradigms on Network Generalization: A Multi-Center IMT Study," in *2023 24th International Conference on Digital Signal Processing (DSP)*, Rhodes (Rodos), Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/DSP58604.2023.10167892.
[138] Z. Sobhaninia *et al.*, "Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany: IEEE, Jul. 2019, pp. 6545–6548. doi: 10.1109/EMBC.2019.8856981.
[139] Y. Fu, J. Chen, and K. Li, "Structure-aware Loss Function for Ultrasound Image Segmentation," in *2021 IEEE International Ultrasonics Symposium (IUS)*, Xi'an, China: IEEE, Sep. 2021, pp. 1–4. doi: 10.1109/IUS52206.2021.9593773.
[140] Y. Xie, H. Liao, D. Zhang, and F. Chen, "Uncertainty-aware Cascade Network for Ultrasound Image Segmentation with Ambiguous Boundary," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, vol. 13434, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., in Lecture Notes in Computer Science, vol. 13434. , Cham: Springer Nature Switzerland, 2022, pp. 268–278. doi: 10.1007/978-3-031-16440-8_26.
[141] K. Lee, J. Y. Kim, M. H. Lee, C.-H. Choi, and J. Y. Hwang, "Imbalanced Loss-Integrated Deep-Learning-Based Ultrasound Image Analysis for Diagnosis of Rotator-Cuff Tear," *Sensors*, vol. 21, no. 6, p. 2214, Mar. 2021, doi: 10.3390/s21062214.
[142] Y. Gong *et al.*, "SCCNet: Self-correction boundary preservation with a dynamic class prior filter for high-variability ultrasound image segmentation," *Comput. Med. Imaging Graph.*, vol. 104, p. 102183, Mar. 2023, doi: 10.1016/j.compmedimag.2023.102183.
[143] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound Image Segmentation: A Deeply Supervised Network With Attention to Boundaries," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1637–1648, Jun. 2019, doi: 10.1109/TBME.2018.2877577.
[144] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified Perceptual Parsing for Scene Understanding," 2018, doi: 10.48550/ARXIV.1807.10221.
[145] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, doi: 10.48550/ARXIV.2201.03545.
[146] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2017, doi: 10.48550/ARXIV.1708.02002.
[147] A. Celaya, B. Riviere, and D. Fuentes, "A Generalized Surface Loss for Reducing the Hausdorff Distance in Medical Imaging Segmentation," 2023, doi: 10.48550/ARXIV.2302.03868.
[148] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," 2018, doi: 10.48550/ARXIV.1812.07032.
[149] P. Lochner *et al.*, "Optic nerve sheath diameter: present and future perspectives for neurologists and critical care physicians," *Neurol. Sci.*, vol. 40, no. 12, pp. 2447–2457, Dec. 2019, doi: 10.1007/s10072-019-04015-x.
[150] K. Helmke and H. C. Hansen, "Fundamentals of transorbital sonographic evaluation of optic nerve sheath expansion under intracranial hypertension: II. Patient study," *Pediatr. Radiol.*, vol. 26, no. 10, pp. 706–710, Oct. 1996, doi: 10.1007/BF01383384.
[151] D. Liu and M. Kahn, "Measurement and Relationship of Subarachnoid Pressure of the Optic Nerve to Intracranial Pressures in Fresh Cadavers," *Am. J. Ophthalmol.*, vol. 116, no. 5, pp. 548–556, Nov. 1993, doi: 10.1016/S0002-9394(14)73195-2.
[152] C. Robba *et al.*, "Ultrasound non-invasive measurement of intracranial pressure in neurointensive care: A prospective observational study," *PLOS Med.*, vol. 14, no. 7, p. e1002356, Jul. 2017, doi: 10.1371/journal.pmed.1002356.
[153] R. Moretti, B. Pizzi, F. Cassini, and N. Vivaldi, "Reliability of Optic Nerve Ultrasound for the Evaluation of Patients with Spontaneous Intracranial Hemorrhage," *Neurocrit. Care*, vol. 11, no. 3, pp. 406–410, Dec. 2009, doi: 10.1007/s12028-009-9250-8.
[154] T. Geeraerts *et al.*, "Ultrasonography of the optic nerve sheath may be useful for detecting raised intracranial pressure after severe brain injury," *Intensive Care Med.*, vol. 33, no. 10, pp. 1704–1711, Sep. 2007, doi: 10.1007/s00134-007-0797-6.
[155] A. Sallam *et al.*, "The Diagnostic Accuracy of Noninvasive Methods to Measure the Intracranial Pressure: A Systematic Review and Meta-analysis," *Anesth. Analg.*, vol. 132, no. 3, pp. 686–695, Mar. 2021, doi: 10.1213/ANE.0000000000005189.

[156] I. Soliman *et al.*, "New Optic Nerve Sonography Quality Criteria in the Diagnostic Evaluation of Traumatic Brain Injury," *Crit. Care Res. Pract.*, vol. 2018, p. e3589762, Apr. 2018, doi: 10.1155/2018/3589762.
[157] P. Lochner *et al.*, "B-Mode Transorbital Ultrasonography for the Diagnosis of Idiopathic Intracranial Hypertension: A Systematic Review and Meta-Analysis," *Ultraschall Med. - Eur. J. Ultrasound*, vol. 40, no. 02, pp. 247–252, Apr. 2019, doi: 10.1055/a-0719-4903.
[158] P. Lochner *et al.*, "Sonography of optic nerve sheath diameter identifies patients with middle cerebral artery infarction at risk of a malignant course: a pilot prospective observational study," *J. Neurol.*, vol. 267, no. 9, pp. 2713–2720, Sep. 2020, doi: 10.1007/s00415-020-09906-0.
[159] C. Schroeder *et al.*, "Atrophy of optic nerve detected by transorbital sonography in patients with demyelinating diseases of the central nervous system," *Eur. J. Neurol.*, vol. 27, no. 4, pp. 626–632, Apr. 2020, doi: 10.1111/ene.14137.
[160] G. A. Stead, F. V. Cresswell, S. Jjunju, P. K. N. Oanh, G. E. Thwaites, and J. Donovan, "The role of optic nerve sheath diameter ultrasound in brain infection," *eNeurologicalSci*, vol. 23, p. 100330, Jun. 2021, doi: 10.1016/j.ensci.2021.100330.
[161] N. Carraro, G. Servillo, V. Maria Sarra, A. Bignamini, G. Pizzolato, and M. Zorzon, "Optic Nerve and its Arterial-Venous Vascularization: An Ultrasonologic Study in Multiple Sclerosis Patients and Healthy Controls," *J. Neuroimaging*, vol. 24, no. 3, pp. 273–277, May 2014, doi: 10.1111/j.1552-6569.2012.00758.x.
[162] G. G. R. J. Johnson *et al.*, "Accuracy of Optic Nerve Sheath Diameter Measurements in Pocket-Sized Ultrasound Devices in a Simulation Model," *Front. Med.*, vol. 9, 2022, Accessed: Oct. 12, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2022.831778
[163] R. Soroushmehr *et al.*, "Automated Optic Nerve Sheath Diameter Measurement Using Super-pixel Analysis," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany: IEEE, Jul. 2019, pp. 2793–2796. doi: 10.1109/EMBC.2019.8856449.
[164] K. M. Meiburger *et al.*, "Automatic Optic Nerve Measurement: A New Tool to Standardize Optic Nerve Assessment in Ultrasound B-Mode Images," *Ultrasound Med. Biol.*, vol. 46, no. 6, pp. 1533–1544, Jun. 2020, doi: 10.1016/j.ultrasmedbio.2020.01.034.
[165] R. R. F. Stevens *et al.*, "An Automated Algorithm for Optic Nerve Sheath Diameter Assessment from B-mode Ultrasound Images," *J. Neuroimaging*, vol. 31, no. 4, pp. 724–732, Jul. 2021, doi: 10.1111/jon.12851.
[166] K. M. Meiburger, A. Naldi, P. Lochner, and F. Marzola, "Automatic segmentation of the optic nerve in transorbital ultrasound images using a deep learning approach," in *2021 IEEE International Ultrasonics Symposium (IUS)*, Xi'an, China: IEEE, Sep. 2021, pp. 1–4. doi: 10.1109/IUS52206.2021.9593827.
[167] F. Marzola, P. Lochner, A. Naldi, R. Lemor, J. Stögbauer, and K. M. Meiburger, "Development of a Deep Learning–Based System for Optic Nerve Characterization in Transorbital Ultrasound Images on a Multicenter Data Set," *Ultrasound Med. Biol.*, vol. 49, no. 9, pp. 2060–2071, Sep. 2023, doi: 10.1016/j.ultrasmedbio.2023.05.011.
[168] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972, doi: 10.1145/361237.361242.
[169] T. Geeraerts, S. Merceron, D. Benhamou, B. Vigué, and J. Duranteau, "Non-invasive assessment of intracranial pressure using ocular sonography in neurocritical care patients," *Intensive Care Med.*, vol. 34, no. 11, pp. 2062–2067, Nov. 2008, doi: 10.1007/s00134-008-1149-x.
[170] Francesco Marzola, "Dataset for - Development of a Deep Learning-based system for Optic Nerve characterization in Transorbital Ultrasound Images on a multicenter dataset." Mendeley, Jun. 16, 2023. doi: 10.17632/KW8GVP8M8X.1.
[171] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[172] K. Christensen-Jeffries *et al.*, "Super-resolution Ultrasound Imaging," *Ultrasound Med. Biol.*, vol. 46, no. 4, pp. 865–891, Apr. 2020, doi: 10.1016/j.ultrasmedbio.2019.11.013.
[173] J. K. Mah and N. Van Alfen, "Neuromuscular Ultrasound: Clinical Applications and Diagnostic Values," *Can. J. Neurol. Sci. J. Can. Sci. Neurol.*, vol. 45, no. 6, pp. 605–619, Nov. 2018, doi: 10.1017/cjn.2018.314.
[174] I. M. P. Arts *et al.*, "Muscle ultrasonography: A diagnostic tool for amyotrophic lateral sclerosis," *Clin. Neurophysiol.*, vol. 123, no. 8, pp. 1662–1667, Aug. 2012, doi: 10.1016/j.clinph.2011.11.262.
[175] E. M. Strasser, T. Draskovits, M. Praschak, M. Quittan, and A. Graf, "Association between ultrasound measurements of muscle thickness, pennation angle, echogenicity and skeletal muscle strength in the elderly," *AGE*, vol. 35, no. 6, pp. 2377–2388, Dec. 2013, doi: 10.1007/s11357-013-9517-z.
[176] J. Z. Heckmatt, S. Leeman, and V. Dubowitz, "Ultrasound imaging in the diagnosis of muscle disease," *J. Pediatr.*, vol. 101, no. 5, pp. 656–660, Nov. 1982, doi: 10.1016/S0022-3476(82)80286-2.

[177] S. Pillen and N. Van Alfen, “Muscle ultrasound from diagnostic tool to outcome measure—Quantification is the challenge,” *Muscle Nerve*, vol. 52, no. 3, pp. 319–320, Sep. 2015, doi: 10.1002/mus.24613.
[178] J. Nijboer-Oosterveld, N. Van Alfen, and S. Pillen, “New normal values for quantitative muscle ultrasound: Obesity increases muscle echo intensity,” *Muscle Nerve*, vol. 43, no. 1, pp. 142–143, Jan. 2011, doi: 10.1002/mus.21866.
[179] C. Sahinis and E. Kellis, “Hamstring Muscle Quality Properties Using Texture Analysis of Ultrasound Images,” *Ultrasound Med. Biol.*, vol. 49, no. 2, pp. 431–440, Feb. 2023, doi: 10.1016/j.ultrasmedbio.2022.09.011.
[180] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[181] M. M. Galloway, “Texture analysis using gray level run lengths,” *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975, doi: 10.1016/S0146-664X(75)80008-6.
[182] F. Molinari, C. Caresio, U. R. Acharya, M. R. K. Mookiah, and M. A. Minetto, “Advances in Quantitative Muscle Ultrasonography Using Texture Analysis of Ultrasound Images,” *Ultrasound Med. Biol.*, vol. 41, no. 9, pp. 2520–2532, Sep. 2015, doi: 10.1016/j.ultrasmedbio.2015.04.021.
[183] T. Watanabe *et al.*, “Quantitative Sonographic Assessment of the Quadriceps Femoris Muscle in Healthy Japanese Adults: Quantitative Sonographic Assessment of the Quadriceps Femoris Muscle,” *J. Ultrasound Med.*, vol. 36, no. 7, pp. 1383–1395, Jul. 2017, doi: 10.7863/ultra.16.07054.
[184] J. J. Martínez-Payá, J. Ríos-Díaz, M. E. Del Baño-Aledo, J. I. Tembl-Ferrairó, J. F. Vazquez-Costa, and F. Medina-Mirapeix, “Quantitative Muscle Ultrasonography Using Textural Analysis in Amyotrophic Lateral Sclerosis,” *Ultrason. Imaging*, vol. 39, no. 6, pp. 357–368, Nov. 2017, doi: 10.1177/0161734617711370.
[185] J. J. Martínez-Payá, J. Ríos-Díaz, F. Medina-Mirapeix, J. F. Vázquez-Costa, and M. E. Del Baño-Aledo, “Monitoring Progression of Amyotrophic Lateral Sclerosis Using Ultrasound Morpho-Textural Muscle Biomarkers: A Pilot Study,” *Ultrasound Med. Biol.*, vol. 44, no. 1, pp. 102–109, Jan. 2018, doi: 10.1016/j.ultrasmedbio.2017.09.013.
[186] J. S. Song *et al.*, “The Relationship Between Muscle Size and Strength Does not Depend on Echo Intensity in Healthy Young Adults,” *J. Clin. Densitom.*, vol. 24, no. 3, pp. 406–413, Jul. 2021, doi: 10.1016/j.jocd.2020.09.002.
[187] P.-T. Liu, T.-S. Wei, and C. T.-S. Ching, “Quantitative Ultrasound Texture Analysis to Assess the Spastic Muscles in Stroke Patients,” *Appl. Sci.*, vol. 11, no. 1, p. 11, Dec. 2020, doi: 10.3390/app11010011.
[188] F. Marzola, N. Van Alfen, J. Doorduin, and K. M. Meiburger, “Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment,” *Comput. Biol. Med.*, vol. 135, p. 104623, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104623.
[189] F. Marzola, N. V. Alfen, M. Salvi, B. D. Santi, J. Doorduin, and K. M. Meiburger, “Automatic segmentation of ultrasound images of gastrocnemius medialis with different echogenicity levels using convolutional neural networks,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada: IEEE, Jul. 2020, pp. 2113–2116. doi: 10.1109/EMBC44109.2020.9176343.
[190] R. J. Cunningham, P. J. Harding, and I. D. Loram, “Real-Time Ultrasound Segmentation, Analysis and Visualisation of Deep Cervical Muscle Structure,” *IEEE Trans. Med. Imaging*, vol. 36, no. 2, pp. 653–665, Feb. 2017, doi: 10.1109/TMI.2016.2623819.
[191] G.-Q. Zhou and Y.-P. Zheng, “Automatic Fascicle Length Estimation on Muscle Ultrasound Images With an Orientation-Sensitive Segmentation,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 12, pp. 2828–2836, Dec. 2015, doi: 10.1109/TBME.2015.2445345.
[192] M. Rana, G. Hamarneh, and J. M. Wakeling, “Automated tracking of muscle fascicle orientation in B-mode ultrasound images,” *J. Biomech.*, vol. 42, no. 13, pp. 2068–2073, Sep. 2009, doi: 10.1016/j.jbiomech.2009.06.003.
[193] C. Caresio, M. Salvi, F. Molinari, K. M. Meiburger, and M. A. Minetto, “Fully Automated Muscle Ultrasound Analysis (MUSA): Robust and Accurate Muscle Thickness Measurement,” *Ultrasound Med. Biol.*, vol. 43, no. 1, pp. 195–205, Jan. 2017, doi: 10.1016/j.ultrasmedbio.2016.08.032.
[194] M. Salvi, C. Caresio, K. M. Meiburger, B. De Santi, F. Molinari, and M. A. Minetto, “Transverse Muscle Ultrasound Analysis (TRAMA): Robust and Accurate Segmentation of Muscle Cross-Sectional Area,” *Ultrasound Med. Biol.*, vol. 45, no. 3, pp. 672–683, Mar. 2019, doi: 10.1016/j.ultrasmedbio.2018.11.012.
[195] G.-Q. Zhou *et al.*, “A Single-Shot Region-Adaptive Network for Myotendinous Junction Segmentation in Muscular Ultrasound Images,” *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 67, no. 12, pp. 2531–2542, Dec. 2020, doi: 10.1109/TUFFC.2020.2979481.
[196] J. Zhou, Y. Lu, S. Tao, X. Cheng, and C. Huang, “E-Res U-Net: An improved U-Net model for segmentation of muscle images,” *Expert Syst. Appl.*, vol. 185, p. 115625, Dec. 2021, doi: 10.1016/j.eswa.2021.115625.

[197] S. Katakis, N. Barotsis, A. Kakotaritis, G. Economou, E. Panagiotopoulos, and G. Panayiotakis, “Automatic Extraction of Muscle Parameters with Attention UNet in Ultrasonography,” *Sensors*, vol. 22, no. 14, p. 5230, Jul. 2022, doi: 10.3390/s22145230.
[198] A. Kirillov *et al.*, “Segment Anything,” 2023, doi: 10.48550/ARXIV.2304.02643.
[199] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: An experimental study,” *Med. Image Anal.*, vol. 89, p. 102918, Oct. 2023, doi: 10.1016/j.media.2023.102918.
[200] W. Zhang, J. Pang, K. Chen, and C. C. Loy, “K-Net: Towards Unified Image Segmentation,” 2021, doi: 10.48550/ARXIV.2106.14855.
[201] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
[202] MMSegmentation Contributors, “MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark.” [Online]. Available: https://github.com/open-mmlab/mmsegmentation
[203] J. J. M. Van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.
[204] G. Thibault *et al.*, “Texture indexes and gray level size zone matrix. Application to cell nuclei classification,” 2009. Accessed: Feb. 19, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Texture-indexes-and-gray-level-size-zone-matrix.-to-Thibault-Fertil/fec6bd9b7f5d6a50410109991857494c8d25f290
[205] C. Sun and W. G. Wee, “Neighboring gray level dependence matrix for texture classification,” *Comput. Graph. Image Process.*, vol. 20, no. 3, p. 297, Nov. 1982, doi: 10.1016/0146-664X(82)90093-4.
[206] M. Amadasun and R. King, “Textural features corresponding to textural properties,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 5, pp. 1264–1274, Oct. 1989, doi: 10.1109/21.44046.
[207] F. Pedregosa, F. Bach, and A. Gramfort, “On the Consistency of Ordinal Regression Methods,” 2014, doi: 10.48550/ARXIV.1408.2327.
[208] A. Nichol and P. Dhariwal, “Improved Denoising Diffusion Probabilistic Models,” 2021, doi: 10.48550/ARXIV.2102.09672.
[209] P. Esser, R. Rombach, and B. Ommer, “Taming Transformers for High-Resolution Image Synthesis,” 2020, doi: 10.48550/ARXIV.2012.09841.
[210] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic Image Synthesis with Spatially-Adaptive Normalization,” 2019, doi: 10.48550/ARXIV.1903.07291.
[211] A. Nichol *et al.*, “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” 2021, doi: 10.48550/ARXIV.2112.10741.
[212] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” 2023, doi: 10.48550/ARXIV.2302.05543.
[213] T. Fujioka *et al.*, “Breast Ultrasound Image Synthesis using Deep Convolutional Generative Adversarial Networks,” *Diagnostics*, vol. 9, no. 4, p. 176, Nov. 2019, doi: 10.3390/diagnostics9040176.
[214] T. Fujioka *et al.*, “Virtual Interpolation Images of Tumor Development and Growth on Breast Ultrasound Image Synthesis With Deep Convolutional Generative Adversarial Networks,” *J. Ultrasound Med.*, vol. 40, no. 1, pp. 61–69, Jan. 2021, doi: 10.1002/jum.15376.
[215] N. J. Cronin, T. Finni, and O. Seynnes, “Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images,” *Comput. Methods Programs Biomed.*, vol. 196, p. 105583, Nov. 2020, doi: 10.1016/j.cmpb.2020.105583.
[216] J. Liang *et al.*, “Synthesis and Edition of Ultrasound Images via Sketch Guided Progressive Growing GANS,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA: IEEE, Apr. 2020, pp. 1793–1797. doi: 10.1109/ISBI45749.2020.9098384.
[217] A. Gilbert, M. Marciniak, C. Rodero, P. Lamata, E. Samset, and K. Mcleod, “Generating Synthetic Labeled Data From Existing Anatomical Models: An Example With Echocardiography Segmentation,” *IEEE Trans. Med. Imaging*, vol. 40, no. 10, pp. 2783–2794, Oct. 2021, doi: 10.1109/TMI.2021.3051806.
[218] C. Tiago *et al.*, “A Data Augmentation Pipeline to Generate Synthetic Labeled Datasets of 3D Echocardiography Images Using a GAN,” *IEEE Access*, vol. 10, pp. 98803–98815, 2022, doi: 10.1109/ACCESS.2022.3207177.
[219] C. Tiago, S. R. Snare, J. Šprem, and K. McLeod, “A Domain Translation Framework With an Adversarial Denoising Diffusion Model to Generate Synthetic Datasets of Echocardiography Images,” *IEEE Access*, vol. 11, pp. 17594–17602, 2023, doi: 10.1109/ACCESS.2023.3246762.

[220] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, “Echo from Noise: Synthetic Ultrasound Image Generation Using Diffusion Models for Real Image Segmentation,” in *Simplifying Medical Ultrasound*, vol. 14337, B. Kainz, A. Noble, J. Schnabel, B. Khanal, J. P. Müller, and T. Day, Eds., in Lecture Notes in Computer Science, vol. 14337. , Cham: Springer Nature Switzerland, 2023, pp. 34–43. doi: 10.1007/978-3-031-44521-7_4.
[221] S. Ogle, A. Merz, R. Parina, M. Alsayed, and M. Milas, “Ultrasound and the Evaluation of Pediatric Thyroid Malignancy: Current Recommendations for Diagnosis and Follow-up,” *J. Ultrasound Med.*, vol. 37, no. 10, pp. 2311–2324, Oct. 2018, doi: 10.1002/jum.14593.
[222] S. S. Ahn, E.-K. Kim, D. R. Kang, S.-K. Lim, J. Y. Kwak, and M. J. Kim, “Biopsy of Thyroid Nodules: Comparison of Three Sets of Guidelines,” *Am. J. Roentgenol.*, vol. 194, no. 1, pp. 31–37, Jan. 2010, doi: 10.2214/AJR.09.2822.
[223] L. R. Remonti, C. K. Kramer, C. B. Leitão, L. C. F. Pinto, and J. L. Gross, “Thyroid Ultrasound Features and Risk of Carcinoma: A Systematic Review and Meta-Analysis of Observational Studies,” *Thyroid*, vol. 25, no. 5, pp. 538–550, May 2015, doi: 10.1089/thy.2014.0353.
[224] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, “You Only Need Adversarial Supervision for Semantic Image Synthesis,” 2020, doi: 10.48550/ARXIV.2012.04781.
[225] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable Fidelity and Diversity Metrics for Generative Models,” 2020, doi: 10.48550/ARXIV.2002.09797.
[226] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, “Assessing Generative Models via Precision and Recall,” 2018, doi: 10.48550/ARXIV.1806.00035.
[227] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved Precision and Recall Metric for Assessing Generative Models,” 2019, doi: 10.48550/ARXIV.1904.06991.
[228] P. A. Humphrey, “Histopathology of Prostate Cancer,” *Cold Spring Harb. Perspect. Med.*, vol. 7, no. 10, p. a030411, Oct. 2017, doi: 10.1101/cshperspect.a030411.
[229] J. I. Epstein *et al.*, “A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score,” *Eur. Urol.*, vol. 69, no. 3, pp. 428–435, Mar. 2016, doi: 10.1016/j.eururo.2015.06.046.
[230] Y. Zhang, J. Zhang, Y. Song, C. Shen, and G. Yang, “Gleason Score Prediction using Deep Learning in Tissue Microarray Image,” 2020, doi: 10.48550/ARXIV.2005.04886.
[231] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.