INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Deep Learning-Based Calcium Scoring of the Aortic Valve Using 3D TEE: Preliminary Study

Rita Seixas Bairros

Master's in Integrated Decision Support Systems

Supervisor:
PhD Tomás Gomes Silva Serpa Brandão, Assistant Professor, ISCTE-IUL

Co-supervisor:
PhD Luís Manuel Nobre de Brito Elvas, Assistant Professor, ISCTE-IUL

September 2024

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Deep Learning-Based Calcium Scoring of the Aortic Valve Using 3D TEE: Preliminary Study

Rita Seixas Bairros

Master's in Integrated Decision Support Systems

Supervisor:
PhD Tomás Gomes Silva Serpa Brandão, Assistant Professor, ISCTE-IUL

Co-supervisor:
PhD Luís Manuel Nobre de Brito Elvas, Assistant Professor, ISCTE-IUL

September 2024

"The important thing is not to stop questioning. Curiosity has its own reason for existence."

Albert Einstein

Acknowledgements

I must start by thanking my supervisors, Tomás Brandão and Luís Elvas, for their guidance, knowledge, and encouragement throughout this process. Without their knowledge and encouragement this thesis could not have been written.

I also would like to express my gratitude to Professor João Carlos Ferreira, the coordinator of this master's program, who has been following my evolution from the start.

I am deeply grateful to Dr. Ana Almeida and Dr. Paula Fazendas, for their collaboration in this project, and the information provided during the sessions we had. I would like to express my gratitude to Dr. Fazendas for her immense patience in collecting the required data and for offering me the opportunity to witness how she approaches this practice directly from the clinical side of this work.

To my parents, the most important people in my life, who have always inspired me with their love and have believed in me through thick and thin. Without them I could not have come this far.

To my boyfriend Ricardo, thank you for your patience, support, and endless encouragement. You have been my rock throughout this journey.

Finally, to my closest friends João, Ricardo and Ana, thank you for your support during these 5 years here in ISCTE, the ups and downs of this academic life.

To all of you, I dedicate this achievement with a deep appreciation and love.

Resumo

Os depósitos de cálcio na válvula aórtica são um fator crucial no diagnóstico da estenose aórtica, uma condição cardiovascular crítica. Nesta dissertação, propõe-se uma abordagem híbrida que combina técnicas de deep learning com processamento de imagem para melhorar a identificação e quantificação das calcificações na válvula aórtica. Foram estabelecidos dois objetivos principais: (1) detetar e extrair a região da imagem correspondente à válvula aórtica, e (2) quantificar os depósitos de cálcio na válvula segmentada, correlacionando os resultados com os scores de Agatston obtidos em TACs. O modelo YOLOv8n foi adaptado para a deteção da válvula, atingindo 99,94% de precisão, 81,82% de recall e mAP de 92,88%. A extração da região de interesse foi bem-sucedida, utilizando segmentação manual e automática. Para quantificação de cálcio, foram exploradas duas abordagens: uma heurística e CNNs, com a ResNet50 ajustada mostrando erro absoluto médio de 1356,56. A precisão do método heurístico foi validada, especialmente em pacientes com scores de cálcio mais elevados, através de uma correlação de Pearson de 0,75 com os scores de Agatston derivados das TACs. Além disso, uma análise com base no género revelou que os pacientes do sexo masculino apresentavam níveis mais elevados de depósitos de cálcio, em linha com estudos anteriores na área cardiovascular. Este trabalho demonstra como a integração de deep learning e técnicas convencionais pode otimizar o diagnóstico da estenose aórtica, contribuindo para diagnósticos mais rápidos e precisos.

Palavras-chave: ecocardiografia; estenose aórtica; quantificação de cálcio; deep learning; processamento de imagem.

Abstract

Aortic stenosis is a critical cardiovascular condition that can be assessed through echocardiography, with calcium deposits on the aortic valve playing a key role in diagnosis. This dissertation presents a hybrid approach combining deep learning and image processing methods to improve the detection and quantification of aortic valve calcifications. Two main objectives were addressed: (1) detecting and extracting the image region corresponding to the aortic valve, and (2) quantifying calcium deposits within the segmented valve, correlating these results with Agatston scores derived from CT scans. An adapted YOLOv8n model was employed for valve detection, achieving 99.94% precision, 81.82% recall, and a mean Average Precision (mAP) of 92.88%. The region of interest was successfully extracted in all cases using a combination of manual annotations and automated segmentation techniques. For calcium scoring, two approaches were explored: a heuristic method and convolutional neural network (CNN) models. The CNN models captured complex patterns in the echocardiographic images, with the fine-tuned ResNet50 model demonstrating superior performance, achieving a mean absolute error of 1356.56. The heuristic method showed a Pearson correlation of 0.75 with the CT-derived Agatston score, validating its accuracy, especially in patients with higher calcium scores. Additionally, a gender-based analysis revealed that male patients exhibited higher calcium deposits, consistent with existing cardiovascular research. This work shows that combining deep learning with traditional methods can improve the diagnostic process for aortic stenosis, offering potential for timely, precise diagnoses and advancing healthcare system efficiency.

Keywords: echocardiography; aortic stenosis; calcium scoring; deep learning; image processing.

Table of contents

List of Figures and Tables

Glossary

CNN – Convolutional Neural Network

CRISP-DM – Cross Industry Standard Process for Data Mining

CT – Computed Tomography

DICOM – Digital Imaging and Communications in Medicine

IoU – Intersection over Union

L2 Regularization – L2 Weight Decay (a regularization technique used to reduce overfitting)

MAE – Mean Absolute Error

mAP – Mean Average Precision

NDA – Non-Disclosure Agreement

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RMSE – Root Mean Squared Error

ROI – Region of Interest

TEE – Transesophageal Echocardiography

U-Net – A deep learning model used for image segmentation

CHAPTER 1

Introduction

Over the last few years, the application of artificial intelligence (AI) to areas such as medical imaging, diagnostic decision support, and predictive analytics in healthcare has transformed various aspects of medical practice. This has contributed to enhanced diagnostic efficiency, more accurate treatment planning, and improved patient care outcomes [1]. AI systems have shown great potential in interpreting large volumes of medical information, assisting doctors in diagnosing diseases at an early stage, and enhancing patients' health [2]. Specifically, the application of AI in medical imaging has been transformative, enabling machines to interpret images in a way that is comparable to human experts or even exceeds in certain domains. In cardiovascular medicine, for instance, AI has enhanced the interpretation of imaging modalities, such as echocardiography, thereby enabling more precise and personalized care for patients [3], [4].

Cardiovascular diseases, and more specifically aortic valve pathologies, are a major global health concern due to their high prevalence and associated morbidity and mortality [5]. Aortic stenosis, characterized by the calcification and narrowing of the aortic valve, requires precise evaluation to guide clinical management. Proper assessment of the severity of aortic valve conditions is critical for effective treatment planning and ongoing monitoring of patients. In a hospital setting, it is especially important to accurately identify patients suffering from aortic valve calcification, as early detection can significantly influence the course of treatment. Automated systems for calcium scoring could streamline this process, helping clinicians prioritize high-risk patients and ensure timely intervention, particularly in cases requiring surgical or interventional procedures. A commonly used method for this assessment is the calcium score of the aortic valve, which quantifies the extent of calcification to reflect disease severity and predict progression [6], [7], [8]. Typically, calcium scores are calculated through computed tomography (CT) scans, which quantify calcium deposits within the aortic valve.

Despite the accuracy of CT-based calcium scoring, this method has several limitations. The most notable drawback is the exposure to ionizing radiation, which poses long-term health risks, especially for patients requiring frequent monitoring [6], [7]. The need to balance diagnostic accuracy and patient safety is particularly pressing in cases where repeated scans are necessary for ongoing assessment. Further complicating the widespread use of CT-based scoring are issues of cost and accessibility, particularly in resource-limited settings where advanced imaging modalities like CT are not always available [9], [10]. These factors underscore the need for alternative methods that can provide accurate and safe assessments without the associated risks of CT imaging.

An alternative to CT is echocardiography, a widely used imaging modality for cardiovascular assessments due to its non-invasive nature, affordability, and diagnostic accuracy [11]. However, the accuracy of echocardiographic assessments can be influenced by the patient's acoustic window, the area through which ultrasound waves pass. Factors such as obesity or lung disease may limit image quality in certain cases [12]. Transesophageal echocardiography (TEE), and more specifically, three-dimensional transesophageal echocardiography (3D TEE), has emerged as a promising tool for visualizing the aortic valve and surrounding structures. 3D TEE offers high-quality, multi-dimensional imaging, making it a suitable candidate for calcium quantification [9], [10], [11]. Unlike CT, 3D TEE is radiation-free, making it an ideal modality for routine follow-ups without exposing patients to ionizing radiation. The detailed anatomical images provided by TEE offer a deeper understanding of valve morphology and function, therefore enhancing clinical decision-making. However, it is important to note that TEE is considered a semi-invasive procedure due to the insertion of an internal probe through the esophagus, which may cause discomfort and requires sedation [13].

TEE circumvents some of these limitations by placing the ultrasound probe closer to the heart, resulting in clearer images and improving diagnostic accuracy, especially in patients with suboptimal transthoracic echocardiographic windows [14].

Recent advancements in deep learning have revolutionized the field of medical imaging, enabling more accurate and efficient analysis of complex medical datasets. Deep learning techniques, particularly convolutional neural networks (CNNs), have been successfully applied to a wide range of medical imaging tasks, including the automatic detection and segmentation of anatomical structures [15]. In cardiovascular imaging, deep learning has shown promise in automating the analysis of echocardiographic data for tasks such as left ventricle segmentation, mitral valve analysis, and disease detection [16], [17].

Building on these advancements, this dissertation proposes the application of deep learning to a new task: automated detection and calcium quantification of the aortic valve using 3D TEE data, with the goal of improving diagnostic accuracy and assisting in the early identification and treatment planning of aortic valve stenosis. Previous research has demonstrated the feasibility of using AI to detect various cardiac conditions based on echocardiographic images, such as mitral regurgitation and pericardial effusion [16], [17]. Extending this technology to the quantification of aortic valve calcification could provide a radiation-free, efficient, and highly accurate solution for the management of aortic stenosis [18].

Moreover, this approach could serve as an effective tool for patient screening, identifying individuals with significant aortic valve calcification. Automating this screening process would assist clinicians in prioritizing patients for further diagnostic evaluation and treatment planning, thus enhancing patient care and reducing reliance on more invasive or resource-intensive procedures such as CT scans [19]. Currently, the manual analysis of each echocardiographic study takes approximately 8 minutes. Given that the hospital handles around 5000 cases annually, this amounts to 666.67 hours or about 28 full working days of physician time. If a physician dedicated three hours daily to this task, it would take approximately 222 days to complete the annual workload. By automating this process, the time could be reduced dramatically, allowing faster patient screening and improving overall clinical workflow efficiency.

This dissertation builds upon recent successes in AI for echocardiographic analysis and addresses the limitations of conventional imaging techniques. By leveraging the high-quality, multi-dimensional data provided by 3D TEE, the deep learning model will automatically detect the aortic valve and quantify calcium deposits, offering a radiation-free alternative to CT-based calcium scoring. Our approach aims to support early patient screening for aortic valve calcification, allowing clinicians to prioritize patients for further diagnostic evaluation and treatment planning.

1.1. Research Objectives and Questions

The primary objective of this dissertation is to automate the assessment of aortic valve stenosis through echocardiographic image. To achieve this objective, this dissertation focuses on two main goals: (1) detecting and extracting the image region corresponding to the aortic valve and (2) quantifying calcium deposits within the segmented valve. By correlating these results with Agatston scores derived from CT scans, this approach aims to streamline the diagnostic process,

reduce reliance on radiation-based methods, and offer clinicians a reliable tool for timely intervention. With the end goal of enhancing patient care, this dissertation also explores how the calcium quantification models could assist in patient screening by identifying individuals with severe calcification, thus aiding in the prioritization of those requiring urgent medical evaluation.

To guide this research, two main research questions were formulated:

RQ1: How accurately can deep learning models, such as YOLOv8, detect the aortic valve in 3D transesophageal echocardiography images?

RQ2: How reliable are the proposed methods for calcium quantification within the segmented aortic valve?

RQ3: How well do the automatically quantified calcium scores correlate with clinically validated scores, such as the Agatston score from CT scans, and aid in patient screening for severe calcification?

1.2. Methodology

This section explains the methodological approach adopted for this dissertation following the CRISP-DM model [20]. The methodology is divided into six phases (Figure 1), each crucial to developing a reliable deep learning model for automated aortic valve calcium quantification. Each phase ensures that data is collected, processed, and analyzed to support the dissertation's objectives. Specific techniques, tools, and processes employed in each phase will be elaborated on in subsequent chapters.

Figure 1 - CRISP-DM Methodology

Business Understanding

Meetings were held with two cardiologists at Hospital Garcia da Orta to gain a deeper understanding of the clinical problem. This collaboration ensured the clinical relevance of the project, with extensive human validation throughout its development.

Data Understanding

Chapter 3 provides an assessment of the dataset collected from Hospital Garcia da Orta, which includes 3D TEE scans focused on the aortic valve. The dataset features key anatomical views that were annotated by cardiologists to assist in the detection and analysis of the aortic valve.

Data Preparation

Data preparation involved applying various data augmentation techniques, such as rotation and zoom, to enhance the training dataset. Additionally, the region of interest was extracted based on expert annotations, ensuring that the dataset remained focused on the aortic valve for optimal machine learning performance. All data preparation steps are described in Chapter 3.

Modeling

Chapter 4 covers the development of various models to detect and quantify calcium deposits. A deep learning model will be used for detection of the aortic valve. For calcium quantification, 2 approaches will be explored, a heuristic method and CNNs.

Evaluation

In Chapter 5, the model’s performance will be evaluated, focusing on its ability to detect and quantify calcium deposits in the aortic valve. The results will be compared to clinical Agatston scores to assess the model's potential clinical relevance and areas for improvement.

Deployment

In the final phase, the research findings will be compiled and presented in this thesis, with a focus on sharing insights with the medical team at Hospital Garcia da Orta. This phase includes documenting the entire process, from data collection and model development to evaluation and conclusions.

CHAPTER 2

State-of-the-art

2.1. Research Methodology

To systematically explore advancements in deep learning applied to aortic valve segmentation and calcium quantification using 3D transesophageal echocardiography, a thorough literature review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [21]. PRISMA provides a structured framework that enhances the transparency and rigor of systematic reviews, allowing for an accurate evaluation and communication of research findings.

The primary research question guiding this review was: "What are the latest advancements in analyzing echocardiography images for determining the calcium score of the aortic valve using deep learning models?"

To ensure comprehensive coverage of the topic, a systematic search strategy was employed. The search process specifically targeted studies addressing key aspects of deep learning, echocardiography, and aortic valve disease. Following the PRISMA framework, the review process included structured planning for the identification of studies, application of inclusion and exclusion criteria, and the critical evaluation of selected papers.

In January 2024, an extensive search was conducted on databases. To optimize the search process, three distinct categories were defined to categorize the search terms:

Methodology: This category focused on techniques used to analyze or process the data.

Image Type: This category centered on the type of data and imaging modality used in the studies.

Context: This category reflected the specific objectives and applications of the research.

The research expression was created by combining components from each category (Methodology, Image Type, Context and Limitations). The query and components for each category are shown in the Table 1 below:

Table 1 - PRISMA Query Elements

Category	Query Elements
Methodology	Computer Vision, Image Classification, Deep Learning
Image Type	Echocardiography, 3D TEE
Context	Aortic Valve Stenosis, Aortic Calcification, Calcium Score
Limitations	Last 5 years. Only articles, reviews and written in English

The search was applied only to the title fields of publications to specifically target studies closely aligned with the research focus. Only reviews and articles written in English were considered for inclusion, and publications within the last five years (2019-2024) were filtered to ensure relevance.

The final search query used was “("Computer Vision" OR "Image Classification" OR "Deep Learning ") AND ("Echocardiography" OR "3D TEE") AND ("Aortic Valve Stenosis" OR "Aortic Calcification" OR "Calcium Score")”.

This query was created to capture the most relevant advancements in the analysis of echocardiography images for determining the calcium score of the aortic valve, with a focus on deep learning, image analysis, and aortic valve disease. In addition to the query results, further literature searches were conducted to explore related fields and techniques that could provide valuable insights and complement the findings, ensuring a comprehensive understanding of the topic.

2.2. Research Results

The search process described earlier yielded 58 results, using Rayyan to assist with screening and duplicate detection [22]. This included 46 papers from Scopus and 12 papers from the Web of Science Core Collection. After removing duplicates, 53 unique papers remained for further evaluation, as shown in Figure 2.

Each of these papers was reviewed by examining their titles and abstracts to assess relevance to the research question and objectives. Studies were excluded if they focused on unrelated populations, imaging modalities, or topics that did not address aortic valve disease, deep learning, or calcium quantification.

Out of the 53 papers, 33 were excluded for not aligning with the study’s scope. Many of these exclusions were due to a focus on cardiac structures other than the aortic valve or the use of traditional image analysis methods without deep learning. After this filtering process, a total of 20 articles were included for detailed analysis in the literature review.

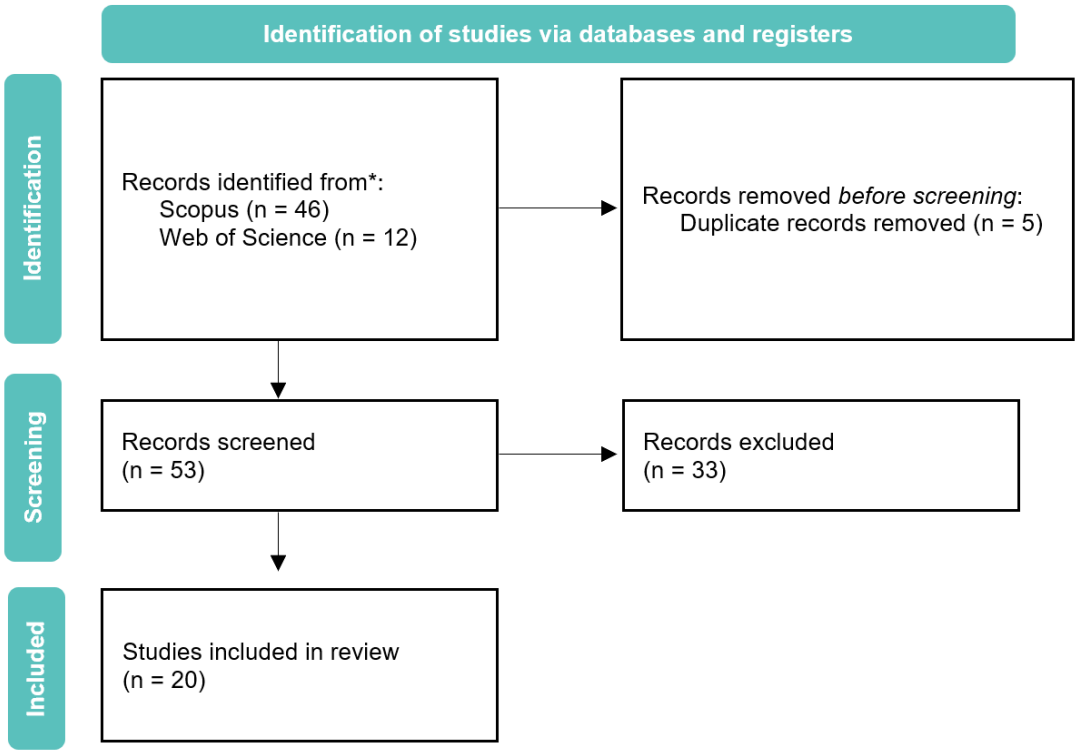

Figure 2 - PRISMA Flow Diagram

The included studies provide a comprehensive view of the ongoing advancements in applying deep learning and AI to echocardiography and other imaging modalities for diagnosing and managing valvular heart diseases (VHDs) and related cardiovascular conditions.

2.3. Review of Retrieved Articles

A key theme in several papers is the automation of critical tasks in echocardiography, particularly for the detection and quantification of valvular anomalies. Nizar et al. [23] and Dai et al. [24] focus on automating the real-time detection of aortic valves and identifying aortic stenosis from parasternal long-axis videos, respectively. These studies highlight the ability of AI models, particularly CNNs, to improve the speed and accuracy of diagnosis during live echocardiographic exams, which could reduce the reliance on human expertise, especially in resource-limited settings. Building on this, Holste et al. [25] developed a deep learning model capable of detecting severe aortic stenosis using 2D echocardiography videos, demonstrating strong generalizability across multiple datasets with high diagnostic performance, thus expanding the applicability of AI to different clinical environments.

Several studies explored AI's role in enhancing the quantification of cardiovascular conditions. Wifstad et al. [26] and Elvas et al. [27] made significant contributions by improving the quantification of valve regurgitation and calcification, respectively, using CNN-based models. Wifstad's approach to measuring regurgitant volume (RVol) from 3D Doppler ultrasound images significantly reduced error rates in orifice and flow volume estimations compared to traditional methods. Elvas and his coauthors work, meanwhile, applied computer vision techniques to semi-automatically quantify calcium in the aortic valve, reducing the need for radiation-heavy CT scans. Similarly, Tang et al. [28] introduced DLFFNet, a model for recognizing aortic valve calcification automatically from echocardiographic images using U-Net for segmentation. This method improved the accuracy of calcification identification, offering an alternative to time-intensive manual assessments.

A key advantage of AI-driven quantification is its ability to minimize variability between observers. Studies like Yang et al. [14], Kim et al. [29]] and Steffner et al. [30] underscore this benefit by demonstrating how AI can automate and improve the efficiency of detecting and quantifying valvular diseases across multiple datasets, even outperforming human experts in some cases. Steffner' work [30] showcases how AI can standardize intraoperative imaging analysis, ensuring consistency across healthcare providers.

AI has also proven valuable in integrating imaging and clinical data. Yuan et al. [31] developed a model for predicting coronary artery calcification using echocardiogram data, demonstrating that AI models can stratify patients' risk comparably to CT-based assessments. Additionally, Karužas et al. [32] explored AI for fully automated aortic measurement in 2D echocardiography, showing that this AI-driven approach not only aligns closely with expert cardiologist measurements but also improves reproducibility. This has important clinical implications for automating routine aortic root assessments, streamlining workflow, and reducing human error.

Beyond echocardiography, AI has been applied to advanced imaging techniques such as 4D flow MRI. Nath et al. [33] developed a deep learning-based network, 4Dflow-VP-Net, for estimating transvalvular pressure gradients noninvasively. This work presents a approach to estimating pressure gradients in stenotic flows, demonstrating high correlation with both catheter-based and Doppler echocardiography measurements. This innovation in MRI-based diagnostics highlights the versatility of AI across different imaging modalities.

Moreover, AI's ability to handle vast amounts of echocardiographic data, particularly multi-view, is another critical advancement. Ahmadi et al. [34] leveraged transformer-based spatio-temporal analysis for classifying aortic stenosis severity. Their deep learning models effectively identified the most informative frames within cine series, avoiding the need for Doppler measurements and expanding the usability of AI tools in clinical practice. This approach reduces the training burden on operators while providing precise, real-time assessments, making AI-based diagnostic tools more accessible to a broader range of clinicians.

Despite these advancements, several studies highlighted challenges that need to be addressed to fully integrate AI into routine clinical practice. Lei et al. [35] and Liu et al. [36] emphasized the importance of training models on large, diverse datasets to ensure their generalizability across different populations and clinical settings. Coulter and Campos [37] further explored real-world applications of AI in echocardiography, acknowledging the efficiency and workflow improvements AI brings but emphasizing that it will complement, rather than replace, the role of physicians in clinical practice.

In summary, the included papers illustrate that AI holds tremendous potential in revolutionizing the field of echocardiography and valvular heart disease management. From improving diagnostic accuracy to automating complex assessments and reducing interobserver variability, AI has shown that it can significantly improve the diagnostic workflow. However, to fully realize these benefits, challenges related to data variability, model generalizability, and clinical integration must be addressed.

2.4. Research Findings

This chapter synthesizes key findings from the systematic review of the selected papers, highlighting advancements in model performance, dataset size, and their contributions to the field.

Table 2 presents a summary of the key findings, model types, and dataset sizes from the included studies. Each paper highlights the application of deep learning and AI for diagnosing VHDs and improving workflow automation.

Table 2 - Key Findings Across AI Models for Valvular Disease and Echocardiography

Paper	Model Type	Dataset Size	Key Findings
[23]	CNN	33 patients	High accuracy in detecting aortic valves using real-time CNN models
[24]	CNN	10 videos, 33 patients	Deep learning for aortic stenosis detection from echocardiograms with high diagnostic performance
[25]	3D CNN	5,257 studies	Strong generalizability of the model for severe aortic stenosis detection across multiple datasets
[26]	CNN	30,000 image pairs	Improved accuracy of regurgitation quantification from Doppler ultrasound
[27]	CV-based calcium quantification	Anonymized patient images	Semi-automatic quantification of calcium deposits, reducing the need for CT scans
[28]	DLFFNet (U-Net)	231 patients	Improved accuracy in AVC identification using local feature fusion approach
[14]	DL framework with CNN-based architecture	1,335 training, 434 test	Automated analysis of valvular heart diseases in Doppler echocardiography with high accuracy
[29]	Contrastive learning	250 video series	Enhanced efficiency in AR diagnosis with multi-view video integration
[30]	CNN	Multi-center dataset	Standardization of TEE view classification, improving intraoperative consistency
[34]	Transformer-based	Public & private datasets	Transformer-based spatio-temporal analysis for AS severity classification
[31]	Video-based CNN	2,881 TTE videos	AI-based CAC prediction comparable to CT scans
[32]	U-Net-based model for automated aortic measurement	58 subjects	Fully automated 2D TTE aortic measurement with close correlation to expert assessment
[33]	4Dflow-VP-Net (DL)	16 patients	High-fidelity pressure gradient estimation from 4D flow MRI

This table summarizes the scope of AI's impact, showcasing improvements in both diagnostic accuracy and workflow automation across a variety of imaging modalities. The studies demonstrate the wide applicability of AI in real-time diagnosis, standardization of processes, and quantitative assessments such as coronary artery calcification (CAC) and pressure gradient estimations.

The AI models presented in the studies utilized various imaging modalities, including 2D and 3D echocardiography, Doppler echocardiography, transesophageal echocardiography (TEE), and 4D flow MRI. While echocardiography (including Doppler and TEE) typically results in 2D or 3D images depending on the acquisition method, modalities such as 4D flow MRI provide more complex datasets, such as 3D point clouds or velocity fields.

Figure 3 above shows the distribution of these imaging modalities across the reviewed studies, highlighting the dominant role of 2D echocardiography and the emerging use of advanced imaging techniques such as 4D flow MRI.

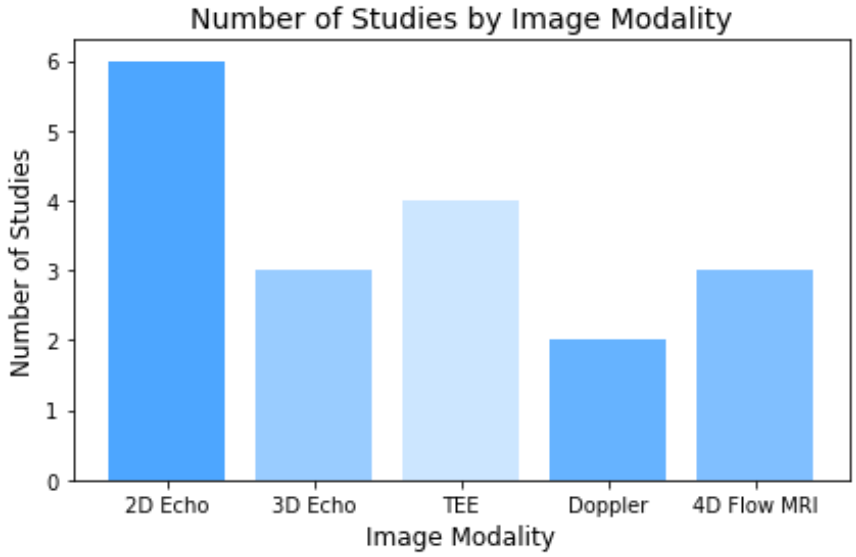

Figure 3 - Number of Studies by Image Modality

Figure 3 indicates that 2D echocardiography is the most widely used imaging modality in the AI-based echocardiography research identified in this review, accounting for 6 studies, followed by TEE and other imaging methods. It reflects the accessibility and established role of 2D echocardiography in routine clinical practice. However, within the broader field of cardiovascular research, other advanced modalities, such as 4D flow MRI and 3D echocardiography, are increasingly being explored, although they remain less represented in the specific studies covered by this review.

In addition to image modality, a key area of comparison is the model performance in specific tasks, particularly in terms of accuracy and AUC (Area Under the Curve). These metrics were used to evaluate tasks such as classifying aortic stenosis severity, detecting aortic valve calcification, and estimating calcium scores. The models analyzed in the selected papers achieve strong performance, with accuracy values close to or exceeding 90% and AUC values nearing 1.0, highlighting the effectiveness of AI models in enhancing diagnostic accuracy for valvular heart diseases.

For example, the models developed by Holste et al. [25] and Tang et al. [28] demonstrates particularly high accuracy, highlighting the value of deep learning techniques like 3D CNNs and U-Nets in automating the detection and quantification of complex conditions like aortic stenosis and valve calcification.

Nevertheless, some models exhibit variation in performance depending on the imaging modality and dataset size. For instance, Steffner et al. [30] reported lower AUC for some TEE view classifications, likely due to the inherent variability in TEE images across different centers. This emphasizes the importance of dataset diversity and the challenges in achieving consistent model performance across diverse clinical environments.

The findings presented in both the table and the figures provide a clear picture of how AI is being effectively integrated into the field of echocardiography. CNNs dominate in terms of model application, showing great success in automating the detection and classification of valvular heart diseases. Transformer-based models and contrastive learning approaches, such as those used by Ahmadi et al. [34] and Kim et al. [29], are gaining traction in improving spatio-temporal analysis and multi-view integration, respectively.

Another important takeaway is the move toward fully automated diagnostic tools that minimize the need for expert intervention while providing highly accurate results. For instance, the automated aortic measurement tool developed by Karužas et al. [32] highlights AI's ability to replace manual measurements with more reproducible, automated assessments.

The use of AI in echocardiography and cardiovascular imaging is expanding, offering significant benefits in terms of diagnostic accuracy, efficiency, and automation. However, while these models have shown excellent performance across controlled datasets, the challenge of generalizability remains. Moving forward, collaboration across medical centers to gather more diverse datasets will be essential for the broader adoption of AI in routine clinical practice. The key studies summarized in Table 2 reflect both the progress made and the future potential of AI-driven echocardiography and cardiovascular care.

2.5. Limitations

Despite the promising advancements highlighted in the reviewed studies, several limitations were consistently noted across the papers and will be outlined in the subsequent subsections. These limitations must be considered carefully, as they affect the real-world applicability and scalability of AI models in clinical settings. Four key challenges were identified from these studies.

Dataset Size

Many of the reviewed papers used small or specialized datasets, limiting the models' generalizability. For example, the study by Nizar et al. [23] utilized a dataset of only 33 patients, which, while sufficient to demonstrate initial model feasibility, is far too limited for broad clinical application. Similarly, studies like Tang et al. [28] and Karužas et al. [32] employed datasets of 231 and 58 patients, respectively, limiting the robustness of the models in handling diverse populations and imaging conditions.

Smaller datasets make it difficult for deep learning models to capture all the inherent variabilities present in real-world clinical data. Variability in patient demographics, imaging quality, and pathological complexity can result in model overfitting to the specific characteristics of the training data. This reduces the model's ability to generalize to new patients or clinical environments. Several papers, including the study by Holste et al. [25], emphasize the need for larger, more diverse multi-center datasets to ensure the generalizability of AI models across different healthcare settings.

Interpretability of AI Models

Another critical limitation frequently raised is the "black box" nature of many AI models, especially deep learning models like CNNs. While models such as those developed by Nizar et al. [23] and Dai et al. [24] show excellent performance in diagnostic tasks, they offer little interpretability for clinicians. This lack of transparency poses a significant barrier to the broader adoption of AI in clinical practice, as physicians must be able to trust and understand the rationale behind a model's predictions.

The need for transparency is particularly crucial in high-stakes environments like cardiology, where treatment decisions directly impact patient outcomes. Tools such as saliency maps, as used by Holste et al. [25], provide some insight into which areas of an image the model is focusing on, but more work is required to bridge the gap between AI predictions and clinical decision-making.

Generalizability Across Different Modalities and Clinical Settings

Several models demonstrated high performance within controlled experimental settings but faced challenges when applied across different imaging modalities or clinical environments. For instance, Steffner et al. [30] highlighted the potential variability in TEE imaging across different centers, which can impact the performance of deep learning models trained in a single institution. The paper by Wifstad et al. [26] also pointed out the limitations of Doppler ultrasound data, which can vary significantly depending on operator skill and equipment quality.

Achieving generalizability across multiple modalities, such as 2D echocardiography, Doppler, and 4D flow MRI, is a key challenge for AI-based models. Models developed and validated using a single modality, such as the video-based CNN for coronary artery calcification prediction in Yuan et al. [31], may not necessarily transfer well to other imaging modalities or clinical workflows.

Clinical Integration and Workflow Impact

While AI offers tremendous potential to improve diagnostic accuracy, the clinical integration of AI tools remains a significant challenge. Studies such as the one by Coulter and Campos [37] discussed the potential for AI to enhance physician workflow rather than replace human expertise. However, the real-world adoption of these tools is often impeded by the need for regulatory approvals, clinician training, and workflow integration.

Besides, the reliance on computational resources and specialized hardware, as noted in Nath et al. [33], presents an additional obstacle in deploying AI models in resource-limited settings. The complexity of training and maintaining deep learning models requires institutional commitment and infrastructure that may not be available in all healthcare systems.

2.6. Research Gaps

While the reviewed studies demonstrate significant advancements in AI applications for echocardiography and VHD management, several critical gaps remain that delay the broader clinical adoption and effectiveness of these technologies.

One of the research gaps in the literature appears to be the relatively limited application of AI models to 3D TEE for calcium quantification in the aortic valve. Although various studies focus on using deep learning for automating the segmentation of cardiac structures and diagnosing valvular diseases [23], [24]. Current methods such as the work by Elvas et al. [27] focus on semi-automated calcium scoring using computer vision techniques with 2D images, often relying on CT for validation. However, the potential of 3D TEE as a non-invasive and radiation-free alternative for calcium scoring is largely unaddressed, leaving a significant gap in the literature.

Despite the promising accuracy rates shown in many studies, there is a notable lack of AI model validation against clinical gold standards like CT-based calcium scoring. For example, while Nath et al. [33] present a deep learning approach to estimate transvalvular pressure gradients from 4D flow MRI, these methods often lack rigorous validation in large-scale, real-world clinical environments. Also, Tang et al. [28] introduced DLFFNet for automatic AVC recognition, but this method lacks comparison with established calcium scoring metrics like the Agatston score from CT imaging.

AI's integration across multiple imaging modalities remains a promising area. Most current research, such as Holste et al. [25] and Steffner et al. [30], focuses on a single modality like 2D or Doppler echocardiography, without exploring how multi-modal approaches could improve diagnostic accuracy. Only a few papers, such as Nath et al. [33], explore multi-modal data integration (4D flow MRI and Doppler). A more comprehensive approach involving multiple modalities (e.g., 3D TEE, 4D flow MRI, and Doppler) would offer a more accurate and holistic assessment of valvular heart disease.

A recurring theme in the limitations is the challenge of ensuring AI model generalizability across diverse clinical environments and patient populations. The small or homogeneous datasets used in many of the reviewed studies (e.g., [23], [32]) limit the broader application of these models in real-world settings. While some authors, attempt to validate their models across multiple datasets [25], issues related to dataset size, diversity, and cross-institutional validation remain underexplored.

While small dataset size remains a challenge in this field, this dissertation focuses on addressing other critical gaps. Specifically, it develops and tests deep learning models for calcium scoring using 3D TEE, offering a radiation-free alternative to CT-based methods. Additionally, it compares these models against clinical standards such as the Agatston score, contributing to the improvement of non-invasive diagnostic tools. However, it is important to acknowledge that the dataset used in this work is also limited in size, and future research with larger datasets will be needed to fully assess the generalizability of these models.

CHAPTER 3

Data Preparation

3.1. Data Understanding

The data used for this dissertation was collected under strict confidentiality agreements to comply with legal and ethical guidelines for protecting personal information. A non-disclosure agreement (NDA) with Hospital Garcia da Orta was also established to ensure the security and confidentiality of patient data. The data collection process was supported by a cardiologist from the hospital, who performed the exams in patients who provided informed consent for the use of their clinical data in this study. These exams were saved in DICOM format, which is the standard format for storing medical imaging data. Each DICOM file contains both the captured images and relevant metadata, such as patient age and sex [38].

To comply with General Data Protection Regulation (GDPR), all personal identifiers were removed through an anonymization process, where filenames were coded (e.g., "case1") to protect patient identities. This ensures that the dataset can be securely used for research while safeguarding patient privacy.

3.2. Dataset

The dataset used in this study includes 154 DICOM images, captured during 3D TEE exams at Hospital Garcia da Orta. These 154 images correspond to 14 individual scans (cases), with each scan consisting of 11 images. Each scan focuses on the aortic valve region and is stored in DICOM format for post-processing. An additional case, consisting of 11 images, was requested to the cardiologist specifically for model validation of the object detection, without a calcium score from CT scans. This brings the total number of images to 165.

In terms of patient demographics, the dataset consists of 14 patients (6 males and 8 females), with an age range from 41 to 88 years. The male patients have a median age of 74.0 years, while the female patients have a median age of 83.5 years. The exams were conducted by the same cardiologist at the hospital, in the context of a pilot study.

During each TEE exam, 3 mm contiguous slices of the aortic valve were obtained, and the region of interest (ROI) was manually traced in all 154 images by the cardiologist using a green line to outline the aortic valve. This ROI includes the valve leaflets and annular calcifications, while non-valvular calcifications in areas such as the left ventricular outflow tract, aortic sinus, coronary arteries, and mitral annulus were excluded from the analysis. CT scans were also conducted within a three-month window before or after the TEE exams, allowing for comparison and validation of the calcium quantification.

All patient data has been anonymized to protect privacy, and a research log has been maintained, linking the anonymized data to a unique case number for future reference and comparison.

3.3. Pre-processing

To prepare the echocardiographic images for analysis, several pre-processing steps were conducted using a combination of OpenCV [39], NumPy [40], Matplotlib [41], and PyDicom libraries [42]. The primary objective of this pre-processing was to utilize the green contour, manually drawn by the medical professional on the echocardiographic images (as shown in Figure 5), which delineates the location of the aortic valve. This contour was crucial for accurately extracting the ROI that contains the aortic valve. By isolating this region, the subsequent analysis focused specifically on the valve, enhancing the accuracy and efficiency of the image processing and machine learning tasks.

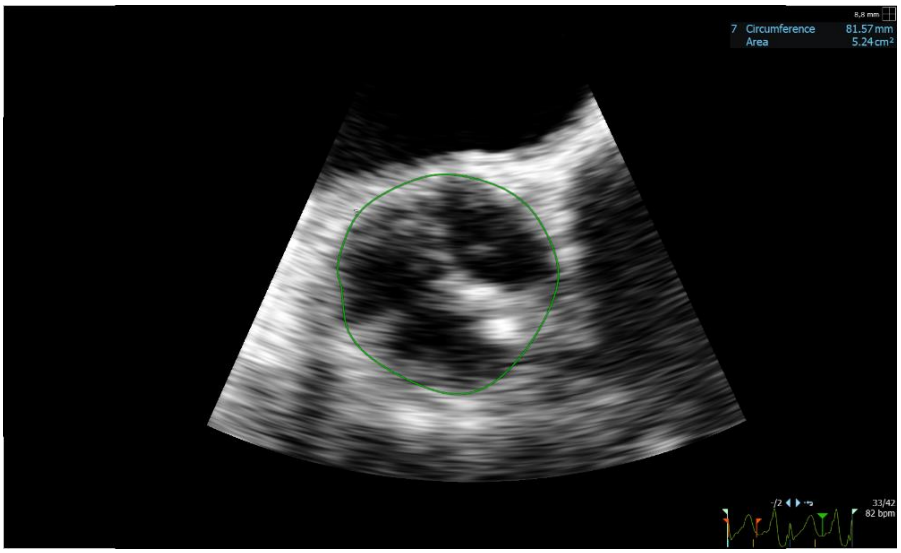

Figure 4 - DICOM with Annotation

3.4. ROI Extraction

The process of extracting the ROI for the aortic valve from echocardiographic images is a critical step in data preparation. This method leverages the green line manually drawn by the medical professional, which demarcates the location of the aortic valve.

The extraction begins by converting the images to the HSV (Hue, Saturation, Value) color space, which is effective for isolating specific colors. A specific range of HSV values corresponding to the green color of the line marking the valve is defined.

Once the HSV values are defined, a binary mask is created that highlights the green areas within the echocardiographic image. This binary mask isolates the green line, allowing the relevant color features to be more effectively identified for the subsequent steps. Morphological operations, such as closing, are then applied to the binary mask to remove small holes and discontinuities, ensuring the detection of continuous and coherent shapes that represent the green line.

Following this, contour detection algorithms are applied to the binary mask to identify the edges of the green line and any enclosed geometric shapes within it. The contours provide valuable information about the structure and shape of the aortic valve. The contour closest in proximity to the green line is identified as the ROI, which corresponds to the aortic valve, as intended by the medical professional.

To finalize the extraction, the green line is removed by drawing a black contour around the identified shape that corresponds to the aortic valve. This step effectively erases the green line, providing a clearer visualization of the valve structure. The final mask is then applied to the original echocardiographic image, isolating the segmented aortic valve without the obstructive green line. This allows for a more accurate analysis and quantification of calcium deposits and other relevant pathological features within the region.

The effectiveness of this segmentation is demonstrated by comparing the original image with the extracted ROI, as demonstrated in Figure 6 below. This segmentation successfully isolated the aortic valve in all cases, demonstrating the precision and reliability of the method, which is essential for subsequent calcium scoring.

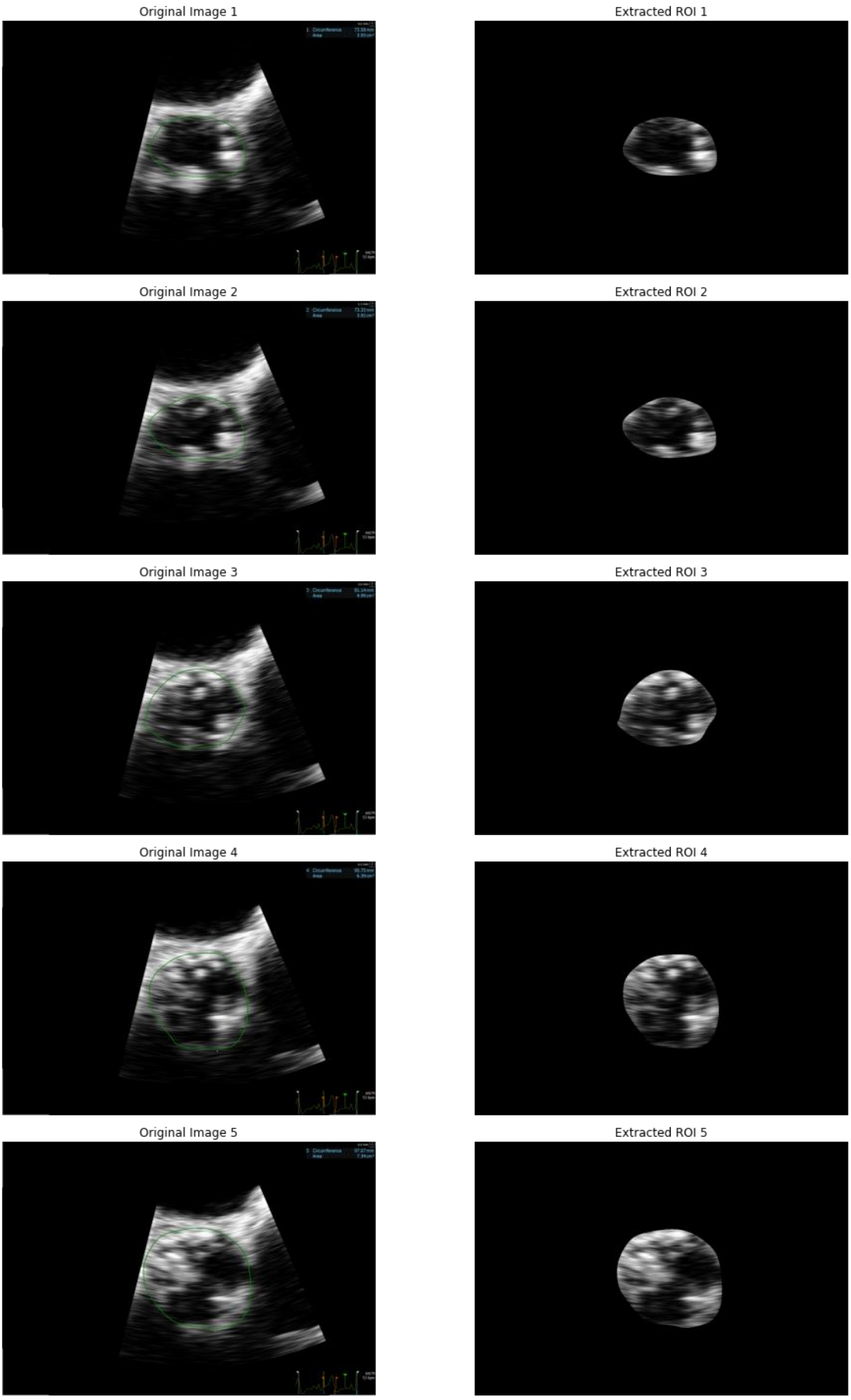

Figure 5 - Comparison of Original Echocardiographic Images and Extracted ROI

3.5. Image Cropping for ROI

As part of the data preparation, it was essential to enhance the quality of the images by cropping them to focus on the ROI. The primary goal of this step was to remove irrelevant areas, such as black regions, while retaining only the essential diagnostic regions that contain calcifications.

This process involved identifying the non-black regions in each image and creating a bounding box around them. The resulting cropped images contained only the relevant portions, ensuring that the dataset was optimized for further analysis. Figure 7 illustrates this process, showing the ROI image on the left and the cropped image on the right.

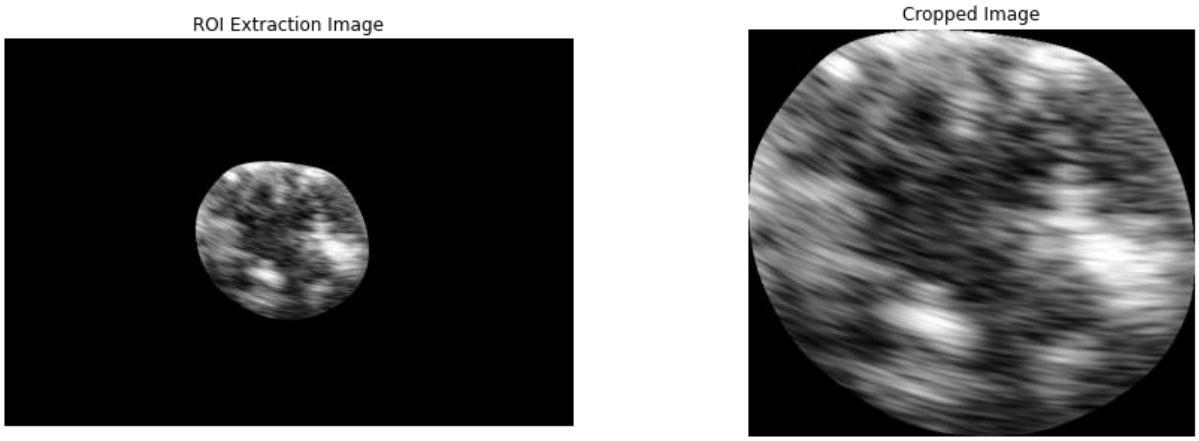

Figure 6 - Comparison of the ROI Image and the Cropped Image

The benefit of this approach is that it reduces noise in the dataset, allowing models to focus on the critical areas for calcium score prediction. Additionally, the cropped images are smaller in size, which helps to improve computational efficiency during the model training process. This focused and clean data improves the overall performance and reliability of the models trained later in the process, including CNNs, by ensuring they work with more relevant input data.

This cropping step is crucial in preparing the dataset for deep learning and machine learning tasks, ensuring that the models can better learn from the key areas of the images and produce more accurate predictions.

3.6. Data Augmentation

Given the limited dataset of 154 echocardiographic images available for training, data augmentation was employed to artificially expand the dataset [43]. This approach is critical for deep learning models, which typically require a large number of diverse training samples to generalize well to unseen data. Data augmentation involves applying a range of transformations to the original images, creating modified versions that expose the model to a broader variety of visual features without the need for additional data collection [44].

The augmentation strategies were implemented as part of the custom configuration used for training the YOLOv8 model and were applied to the original images, not the cropped images. These augmentations were designed to improve the model's ability to detect the aortic valve in various imaging scenarios and were based on guidelines provided by a cardiologist to ensure clinical relevance and realism. The transformations included:

- **Translation:** The images were shifted horizontally and vertically by up to ±20%, simulating slight movements in the imaging procedure.
- **Zoom:** Each image was zoomed in or out by up to 25%, allowing the model to focus on different scales of calcification within the aortic region.
- **Rotation:** The images were rotated by up to ±10 degrees, representing slight variations in patient positioning during image acquisition.
- **Contrast Adjustments:** The brightness and contrast of the images were altered within a range of 0.5 to 1.2, simulating differences in imaging settings or equipment.

Each of these four transformations was applied to each of the 154 original images, resulting in a total of 616 augmented images, ensuring a more diverse and varied set for model training. These augmentations provided the model with a variety of perspectives on each image, improving its robustness by helping it learn to generalize across different scenarios and imaging conditions.

CHAPTER 4

Modeling

The modeling phase of this dissertation is centered on achieving two key objectives as illustrated in Figure 8: (1) the detection and segmentation of the aortic valve in echocardiographic images, and (2) the quantification of calcium deposits within the segmented valve. To accomplish these objectives, a deep learning-based approach is employed, leveraging the YOLOv8 model for object detection. The model is specifically configured to automate the identification and isolation of the aortic valve in echocardiographic images. Once the aortic valve is detected, a post-processing step isolates the valve from the surrounding tissues by applying a circular mask based on the detected bounding box.

Following the detection, the next step involves quantifying the calcium deposits within the valve. Two algorithms are employed for this purpose. One method involves a deep learning-based approach to analyze the ROI and estimate the calcium score. The other is a heuristic approach, where pixel intensity is used as the predictor, and the total count of white pixels within the ROI is used as the calcium score estimate.

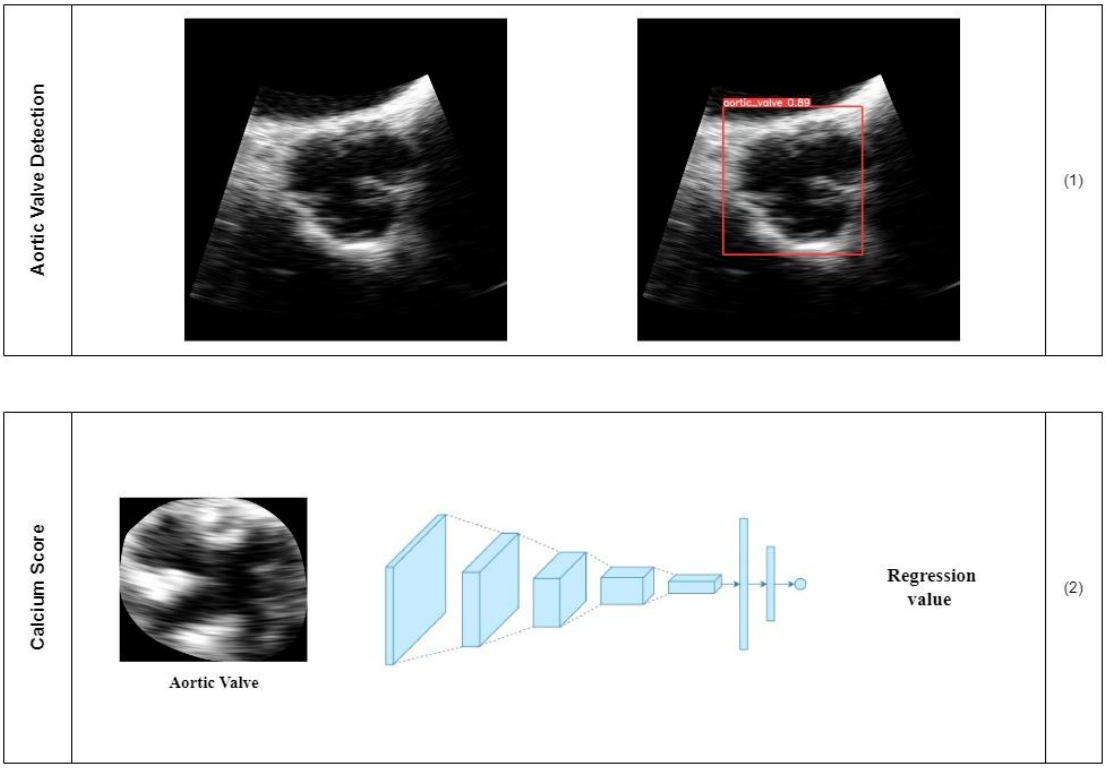

Figure 7 - Framework for Aortic Valve Detection and Calcium Scoring

4.1 Aortic Valve Detection using YOLOv8

The detection of the aortic valve in echocardiographic images is crucial for diagnosing cardiovascular diseases such as aortic stenosis. Accurately identifying the aortic valve is the first step in assessing the valve's condition, including calcification and structural abnormalities. Recently, deep learning models have emerged as a powerful tool to automate this task with high accuracy as seen in the state-of-the-art [45].

This chapter describes the application of YOLOv8n, an object detection model, to detect the aortic valve in echocardiographic images. YOLOv8n, is recognized for its lightweight architecture and rapid inference capabilities, particularly suited for clinical applications that require real-time analysis [46].

By incorporating data augmentation techniques recommended and discussed with a cardiologist, as detailed in Chapter 3, the model was improved to generalize better across various echocardiographic images.

Due to the limited size of the dataset, both the training and validation sets were relatively small. The training set consisted of 154 original images, supplemented by 616 augmented images. For validation, an additional case of 11 images was used, specifically requested to the the cardiologist, as outlined in Chapter 3. Care was taken to ensure that none of the artificially produced data from the training augmentation process overlapped or replicated within the validation set. This ensured the model was truly evaluated on unseen data, maintaining the integrity of the performance assessment. Additionally, none of the validation images included the manually drawn green lines, which were used during the training process for annotation purposes.

A configuration file was also created to specify the dataset paths for images and labels, the number of classes as one, and the class name as aortic valve.

The bounding box annotations for model training were generated by detecting the green circles representing the aortic valve using color-based features. A mask was applied to isolate the green hues, followed by morphological operations to refine the mask. The position and size of the circles were then used to create the bounding boxes.

The model was trained using this dataset for 50 epochs, with a batch size of 16 and an image size of 640x640 pixels. This specific image size was chosen due to the requirements of the YOLOv8 framework, which performs optimally with input dimensions of 640x640. The Adam optimizer was employed, using a learning rate of 0.001. Various augmentation techniques were dynamically applied during training, exposing the model to different transformations in each epoch. The loss function used combined object classification and bounding box regression losses.

To evaluate the model's performance, we used standard object detection metrics, including precision, recall, mean Average Precision (mAP@50), and mAP@50-95. Precision measures the proportion of true positive detections among all positive predictions, while recall measures the proportion of true positives detected out of all actual positives. The mAP@50 score averaged the precision across various recall thresholds at an Intersection over Union (IoU) threshold of 0.5, while mAP@50-95 averaged precision across a range of IoU thresholds (0.5 to 0.95). IoU measures the area of overlap divided by the area of union between the predicted and true bounding boxes, providing a direct measure of how well the model localized the aortic valve.

The YOLOv8n model, selected from the YOLOv8 family due to its lightweight architecture and efficiency, achieved strong performance on the validation dataset, with a mAP@50 of 92.88%, a precision of 99.94%, and a recall of 81.82%. These results highlight the model's capability to detect the aortic valve across different imaging scenarios. However, the model's mAP@50-95 score of 43.55% suggests there is room for improvement when dealing with more stringent IoU thresholds. This metric evaluates the model's performance over a range of IoU thresholds from 0.5 to 0.95, making it a stricter measure of localization accuracy.

Since the inference was performed during model training, the validation dataset results were analyzed to assess the model's performance. The model was able to identify the aortic valve accurately in most of the validation images, with predictions saved as bounding boxes in both image and text formats. Figure 9 provides an example of the model's inference on a validation image, showing the detected aortic valve with the bounding box and confidence score.

Figure 8 - Inference result showing the detected aortic valve in a validation image

For further analysis, a post-processing step was employed, which segmented the detected aortic valve from the background by creating a circular mask based on the bounding box and applying it to the original image. This post-processing step was particularly useful for isolating the aortic valve, allowing for more detailed examination and potentially aiding in further clinical assessments. Figure 10 demonstrates the segmented result, where the aortic valve is clearly isolated, providing a more focused view for clinical analysis.

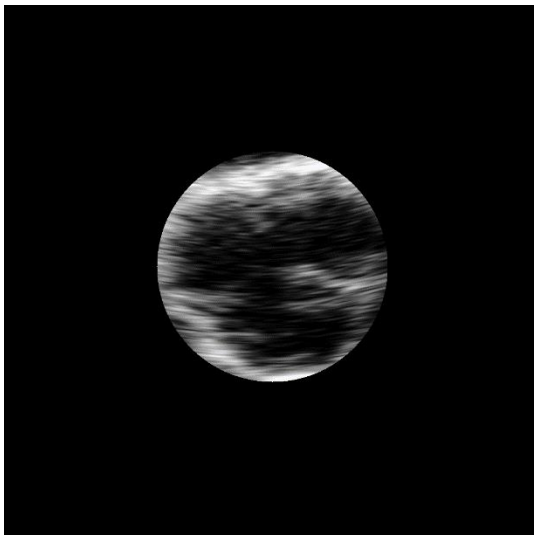

Figure 9 - Segmented result of the aortic valve using the circular mask applied to the bounding box

The inference speed of YOLOv8n was a critical factor for its clinical applicability. On average, the model required 3.72 ms for preprocessing, 191.63 ms for inference, and 3.81 ms for postprocessing per image, with a resolution of (1, 3, 640, 640). Although the inference time could be slightly improved, this processing speed is acceptable for most clinical environments, where accurate detection takes precedence over real-time speed.

Of the 11 validation images, the model successfully detected the aortic valve in 8, with a few images showing no detections or low-confidence predictions. A more in-depth analysis revealed several factors that may have contributed to these missed detections. In some cases, the images had low contrast or excessive noise, which hindered the model's ability to distinguish the boundaries of the aortic valve clearly. Additionally, the presence of large areas of black space in some images likely confused the model, as it relies heavily on visible features to make accurate detections.

The variability in confidence levels among detected valves also highlighted challenges in cases where the valve's visibility was less clear. In images where the valve was more distinct, the model returned high confidence scores (0.7 to 0.9), while in less clear cases, the confidence scores were as low as 0.3 or 0.4. This variation suggests that the model could potentially benefit from additional preprocessing steps, such as enhancing contrast or reducing noise, to improve performance in challenging cases.

Overall, the model's precision, recall, and inference times demonstrate that YOLOv8n holds great promise for automating aortic valve detection in echocardiographic images. However, as seen in the batch of validation results, there are clear areas where performance could be improved, particularly in cases with low contrast or excessive noise. Further refinement of the model, including potential improvements in preprocessing or training with more diverse data, may help to mitigate these challenges and provide even more reliable support for diagnosis and treatment planning.

4.2 Calcium Scoring Approaches

As depicted in Figure 11, we explore two distinct approaches for calculating the calcium score. The first approach is a heuristic method, which involves binarizing the image using a predefined intensity threshold to highlight calcified regions as white pixels. The calcium score is then estimated by summing these white pixels. Afterward, a machine learning model is applied to correlate the estimated calcium score with the CT scores, improving the accuracy of the predictions.

The second approach utilizes CNNs, which are capable of automatically learning complex patterns in medical images and can provide an alternative method for calcium scoring. However, due to the limited number of available images for training, we primarily employed the heuristic method for practical use in the hospital. By leveraging both approaches, we were able to compare and validate the results with the available data.

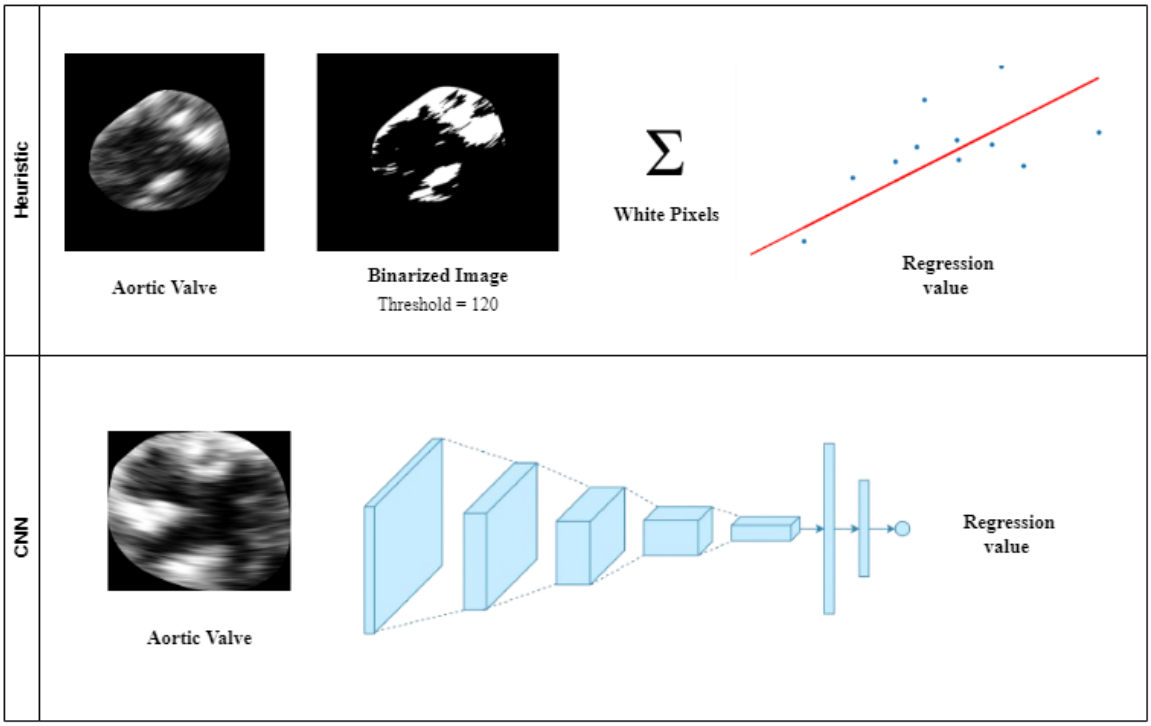

Figure 10 - Calcium Scoring Approaches

4.2.1. Heuristic Method

The heuristic method is a straightforward and computationally efficient approach used in medical imaging for quantifying calcifications, particularly due to its simplicity and computational efficiency [27]. This method works by binarizing the image, converting it into a black-and-white format where the white pixels represent areas of high intensity, which are likely to correspond to calcified regions. The sum of these white pixels provides an estimate of the calcium score, a technique that has been utilized in various cardiovascular imaging studies [47].

The implementation of this method involves several key steps: first, the input image is binarized using an intensity threshold. Pixels with intensity values above this threshold are converted to white, representing potential calcifications, while those below are turned black. The total number of white pixels is then counted, providing a raw estimate of the calcium score. This technique has been widely used in cardiovascular studies to quantify calcification levels [48].

One of the primary advantages of the heuristic method is its simplicity and ease of implementation, making it accessible for various clinical applications, particularly in settings with limited computational resources [27]. However, this threshold is critical in distinguishing true calcifications from other high-intensity artifacts, such as noise or anatomical structures, as noted in threshold-based segmentation studies [49]. Improper threshold selection can lead to underestimation or overestimation of calcifications, affecting the reliability of the calcium score.

Given these potential limitations, the threshold choice could be informed by empirical data and adjusted according to the specific imaging modality and patient population [47]. Furthermore, while the heuristic method may be effective for quick assessments, it can be replaced by more sophisticated techniques, such as machine learning models, to enhance accuracy in complex cases [48]. This leads us to the next subchapter, where alternative methods are explored to address these challenges.

4.2.1.1 Intensity Threshold Selection and Correlation Analysis

A range of intensity thresholds was evaluated to identify the one that most effectively correlates with the actual calcium scores obtained from clinical CT measurements. Different thresholds were systematically tested, and the corresponding calcium scores were calculated. Figure 12 illustrates the correlation coefficients plotted against various intensity thresholds, showing how threshold choice affects the correlation between computed and actual calcium scores.

In this study, three correlation coefficients were employed to assess the relationship between the computed calcium scores and the clinical CT measurements:

- Pearson correlation measures the linear relationship between computed and actual calcium scores. A high Pearson correlation indicates a strong linear relationship, meaning that as the computed score increases, the actual calcium score obtained from CT scans increases proportionally [50].

- Spearman's rank correlation evaluates the strength and direction of a monotonic relationship. Unlike Pearson, Spearman does not assume linearity, making it useful for relationships that are not strictly linear but maintain a consistent ranking [51].
- Kendall's tau assesses the ordinal association between the computed and actual calcium scores, focusing on how well the rankings of the scores match. Kendall's tau is robust in small datasets and noisy data, providing additional reliability [52].

Using these three metrics ensures a comprehensive evaluation of the relationship between the computed calcium scores and clinical measures. Pearson focuses on linear relationships, while Spearman and Kendall offer insight into non-linear and ordinal associations, respectively.

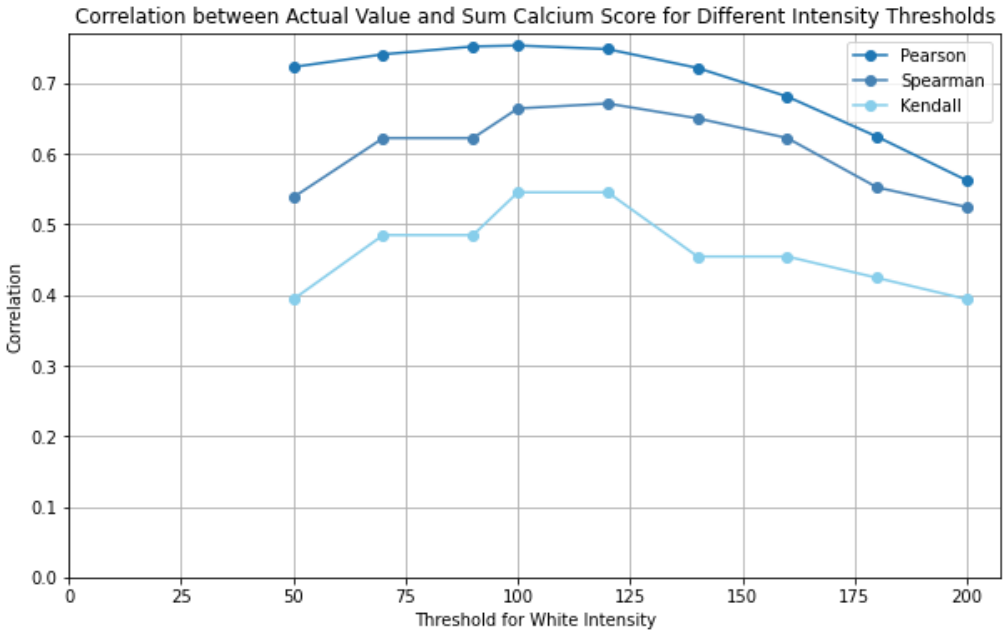

Figure 11 - Correlation Analysis of Intensity Thresholds. Pearson, Spearman, and Kendall correlation coefficients are plotted against varying intensity thresholds.

The results in Figure 12 show that the correlation with clinical calcium scores peaks at a threshold range between 100 and 120 for all metrics. Pearson's correlation coefficient reaches its highest value in this range, indicating a strong linear relationship at these thresholds. Outside of this range, both lower and higher thresholds result in decreased correlations, as lower thresholds introduce more noise by including non-calcified regions, while higher thresholds tend to exclude some relevant calcified areas. Spearman and Kendall correlations exhibit similar trends, showing that the threshold range of 100 to 120 provides the best balance between sensitivity to calcifications and exclusion of irrelevant areas.

Thresholds below 70 resulted in excessive noise, capturing irrelevant high-intensity areas that do not correspond to calcifications. In contrast, thresholds above 130 excluded smaller calcified regions, leading to an underestimation of the calcium score. e threshold range of 70 to 130 provided the best balance between sensitivity to calcifications and exclusion of noise, as demonstrated by the higher correlation values across all metrics.

After identifying the optimal threshold range through correlation analysis, it was important to assess whether these results align with clinical perspectives. This was done by generating binarized images based on the identified thresholds and seeking the opinion of cardiologists for human validation, as discussed in the next section.

4.2.1.2 Expert Validation with Binarized Images

To validate the effectiveness of the selected thresholds, binarized images were generated using the identified threshold range. These images, shown in Figure 13, were reviewed by two cardiologists. The experts provided qualitative feedback on the clarity and accuracy of the binarized images in representing calcified regions.

The cardiologists confirmed that the binarized images provided clear visual representations of potential calcifications, particularly in areas corresponding to known calcified regions on clinical CT scans. Specifically, the threshold of 100 was highlighted as providing the best balance between sensitivity to calcifications and the exclusion of noise, making it clinically relevant for calcification detection.

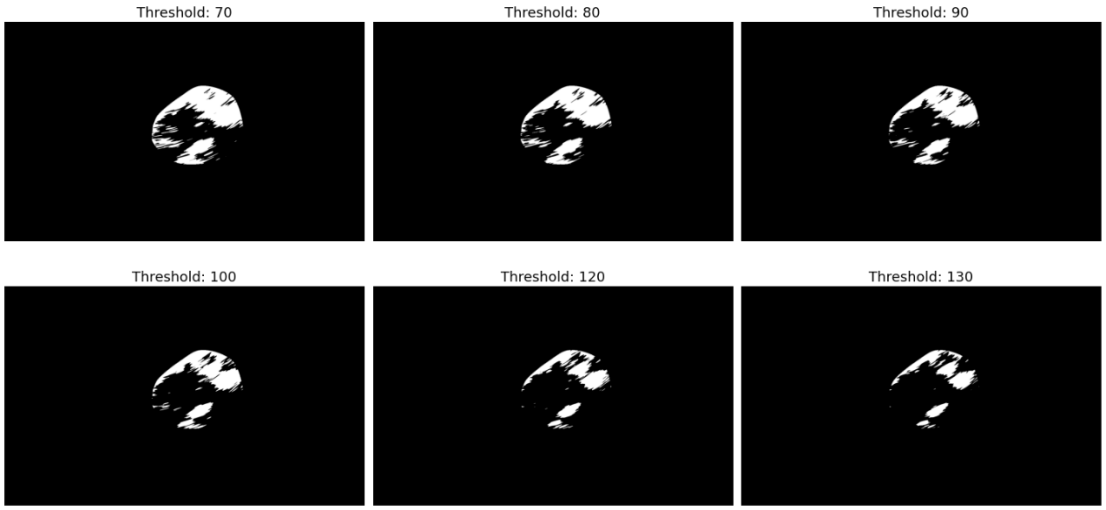

Figure 12 - Binarized Images for Intensity Thresholds (70-130)

This expert validation was crucial in refining the threshold selection process. While the correlation analysis provided a solid foundation based on statistical metrics, expert feedback ensured that the selected thresholds were practically useful in a clinical context. By combining quantitative analysis with expert review, we ensured that the image processing techniques were both technically sound and clinically applicable.

4.2.1.3 CT Calcium Score vs. Predicted Calcium Score

After validating the binarization technique, the analysis progressed comparing the predicted calcium scores obtained through the heuristic method with the actual scores derived from clinical CT scans. The heuristic prediction was calculated as the sum of white pixels across the 11 images corresponding to each case. The scatter plot in Figure 14 illustrates this relationship, where the x-axis represents the CT calcium score, and the y-axis shows the normalized predicted calcium score. Each of the 14 points in the plot corresponds to a case within the dataset, highlighting the comparison between predicted and actual values.

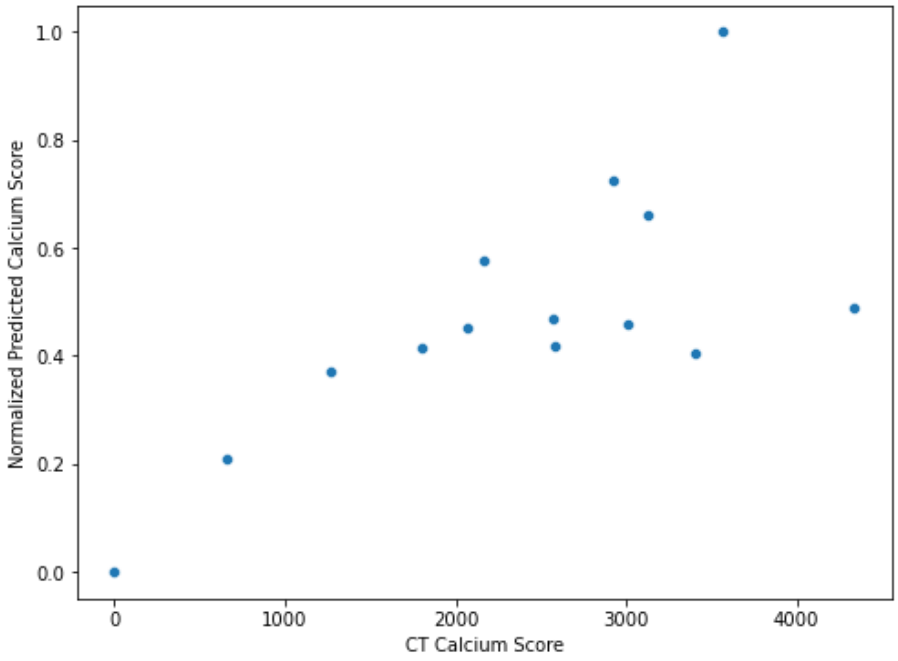

Figure 13 - CT Calcium Score vs Heuristic Calcium Score

The scatter plot reveals a clear positive correlation between the predicted and actual calcium scores. As the sum of white pixels increases, there is a corresponding increase in the CT score, indicating that the heuristic method provides a reasonably accurate estimate of calcification levels.

The Pearson correlation coefficient of 0.75 reinforces this observation, highlighting a strong linear relationship between the predicted and actual scores. While the heuristic method effectively captures the extent of calcification in most cases, there is some variability, likely due to noise in the imaging data or the method's limitations. Despite this, the heuristic approach proves to be a viable tool for estimating calcification, particularly in patients with higher calcium scores in CT scans.

4.2.1.4 Gender-Based Analysis of Calcium Scores

Following the validation of the heuristic method, we conducted further analysis to investigate potential gender-based differences in calcium scores. Figure 15 presents a scatter plot that distinguishes between male and female patients. The x-axis represents the actual calcium score obtained from CT scans, while the y-axis shows the predicted calcium score using the heuristic method. Male patients are represented by blue dots, and female patients by pink dots.

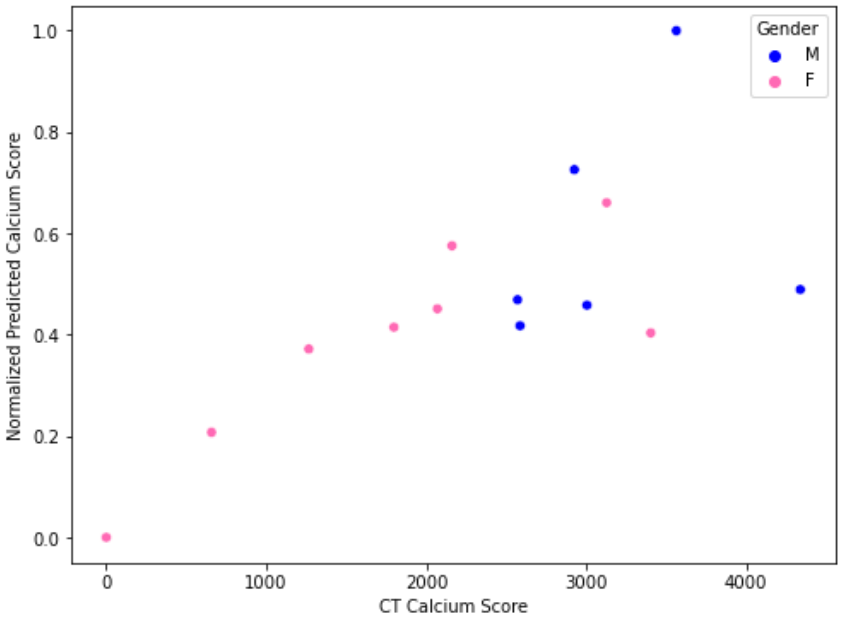

Figure 14 - Calcium Scores by Gender

Despite the small and non-random sample, the plot suggests that male patients tend to have higher calcium scores compared to female patients, both in the predicted and actual scores. This observation aligns with existing cardiovascular research, which indicates that men typically present higher levels of vascular calcification, particularly in conditions like aortic stenosis and other cardiovascular diseases [53], [54].

In the upper-right corner of the plot, several male patients display both high predicted and actual calcium scores, with values exceeding 3000, reinforcing the understanding that men are more prone to severe calcification. This clustering of male patients with severe calcification supports the notion that men are more prone to developing calcific aortic valve disease (CAVD) at a faster rate [55], [56].

Conversely, female patients tend to cluster in the lower and mid-range calcium scores. Although some women show moderately high calcium scores (between 1500 and 2000), most female patients exhibit lower levels of calcification. This pattern is consistent with research suggesting that women tend to develop more diffuse, less localized calcifications, often leading to lower overall calcium scores [57].

The observed gender-based differences have important clinical implications. Men with higher calcium scores are at greater risk of complications related to aortic stenosis and coronary artery disease, supporting the need for earlier and more aggressive interventions in male patients, as they are more likely to experience severe calcification and its associated risks [55].

In contrast, the clustering of lower calcium scores in female patients may indicate a need for gender-specific approaches to risk assessment and treatment. Given the more diffuse nature of calcification in women, alternative imaging methods or more sensitive techniques may be required to fully capture the extent of calcification in female patients [56].

4.2.1.5 Machine Learning Models

Building upon the results obtained with the heuristic method, we evaluated a set of machine learning models to further refine the calcium score predictions. The goal was to improve prediction accuracy by incorporating additional demographic features, such as gender and age, along with the sum of white pixels. The target variable for these models was the CT calcium score, which varied between 0 and 5000 within the dataset. Predictor variables included the sum of white pixels and demographic factors like age and gender.

Given the small dataset of only 14 patients (14 data points) , simple models were chosen to avoid overfitting and ensure that the models remained interpretable. Linear Regression, Lasso Regression, and Ridge Regression were selected as strong baseline, while also being less prone to overfitting in small datasets [58]. More complex models, such as Random Forest and Gradient Boosting, were included to explore potential non-linear relationships. However, due to the limited data, these more complex models did not perform well [59].

The models were evaluated using two primary metrics. Root Mean Squared Error (RMSE) was used to measure prediction error, where lower values indicated better accuracy [60]. R^2 Score was employed to assess how well the model explains the variance in the target variable, with higher values indicating stronger predictive performance [61]. The dataset was split into 80% training and 20% test sets to assess model performance on unseen data.

Table 3 - Performance Metrics of Regression Models for Calcium Score Prediction

Model	RMSE	R^2
Linear Regression	634.6	0.27
Ridge Regression	634.6	0.27
Lasso Regression	635.2	0.27
Random Forest	1081.7	-1.13
Gradient Boosting	1421.0	-2.68

The results summarized in Table 3 provide some level of reliability when compared with the Agatston scores, commonly used as a clinical benchmark for quantifying coronary artery calcification. Although an R^2 of 0.27 indicates that only 27% of the variability in the data is explained by the model, the RMSE value of 634.6 suggests the model's predictions are moderately accurate within this specific dataset.

The RMSE should be interpreted in the context of calcium scores, which in clinical practice can vary from 0 up to several thousand. In this dataset, scores ranged from approximately 0 to 5000, making an RMSE of 634.6 relatively acceptable in cases of high calcification. For higher Agatston scores (e.g., above 3000), an error of this magnitude is proportionally smaller, implying that the model's predictions align reasonably well with true values in severe cases. This means that, despite the model explaining only a fraction of the data variability (as reflected by the R^2), the magnitude of error in higher calcium scores becomes more tolerable, suggesting the model's utility in cases where calcification is significant.

The Lasso and Ridge Regression models showed minimal improvement over Linear Regression, indicating that regularization techniques offered limited benefits due to the dataset's size and the absence of multicollinearity among predictors. The poorer performance of Random Forest and Gradient Boosting models, with RMSEs exceeding 1000 and negative R^2 values, highlights that these more complex models could not effectively generalize due to the limited sample size, reaffirming the reliability of the simpler linear approaches within the context of this analysis.

Interestingly, a further experiment was conducted by including gender and age as features in the models. This led to a deterioration in performance across all metrics, with results notably worse than those observed in the primary models without these features. This could be due to the small dataset, where the inclusion of additional variables might introduce more noise than insight, a phenomenon previously observed in small-sample regression models [6]. It appears that the sum of white pixels alone is the most reliable predictor of calcification in this sample, without the need for demographic adjustments.

The scatter plot in Figure 16 compares the predicted calcium scores with the actual CT-derived scores for the best-performing model (Linear Regression), demonstrating a reasonably linear trend. The red dashed line marks the ideal situation where predicted scores would perfectly align with actual CT scores. While deviations can be observed across the range, errors tend to be more pronounced for higher calcium scores. This indicates that, despite some alignment at the upper end, prediction errors increase as the actual scores rise, challenging the model's consistency and reliability in cases of more severe calcification.

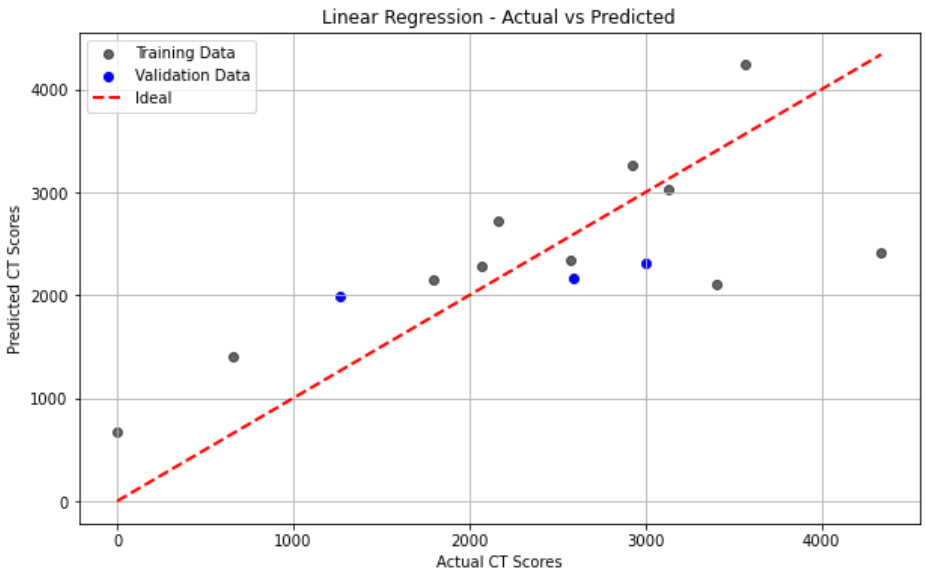

Figure 15 - Scatter Plot of Predicted vs. Actual Calcium Scores (Linear Regression)

4.2.2 CNNs for Calcium Score Prediction

The section on CNNs for calcium score prediction provides an expansion on the prior heuristic and machine learning approaches.

The experiments used a dataset of 154 cropped echocardiographic images, as outlined in Chapter 3, which were specifically prepared to highlight regions of interest where calcifications were most likely. This approach aimed to enhance the model's ability to focus on clinically relevant areas.

The images were uniformly resized to 128x128 pixels and normalized by scaling the pixel values between 0 and 1. The decision to use 128x128 dimensions was based on a balance between computational efficiency and maintaining sufficient resolution for the model to quantify calcifications. This size was chosen after preliminary testing, which indicated that increasing the resolution to 256x256 from the original images did not lead to improved performance.

In subsequent experiments, higher-resolution resizing (such as 256x256 or 224x224) was always performed using the original cropped images, not the already resized 128x128 versions. These larger dimensions were tested to explore whether a finer resolution could improve the model's ability to capture intricate calcification patterns. However, results with 256x256 yielded slightly worse performance, likely due to overfitting and increased model complexity.

The dataset was split into training and validation sets, with 80% of the images used for training and 20% for validation.

Initial CNN Model

The first CNN model consisted of multiple convolutional layers, followed by max-pooling layers to progressively extract features at different levels of granularity. This model aimed to capture patterns related to calcification in echocardiographic images. To improve prediction accuracy, the architecture was specifically designed to handle 11 images per case, providing the model with multiple perspectives of each patient's heart.

Each of the 11 images has an input shape of (128, 128, 3), which corresponds to the image dimensions and color channels. The model processes these 11 separate input layers, where each layer corresponds to a different image from the same patient. This allowed the CNN to learn and extract features from multiple angles of the same anatomical region.

The architecture featured four convolutional layers, each followed by ReLU activations and max-pooling layers. The convolutional layers used 3x3 filters to capture localized features such as edges and textures within the images. After feature extraction, the extracted features from the 11 images were concatenated to aggregate the information across all views of the patient's heart.

Once the features were aggregated, the model applied a fully connected layer with 200 units, followed by an output layer with one unit representing the predicted calcium score. The model was trained using the Adam optimizer with Mean Squared Error (MSE) as the loss function to minimize the prediction error [62].

In terms of results, the model achieved a test loss of 1021.416 and a mean absolute error (MAE) of 726.71 after 10 training epochs. While the MAE of 726 indicated reasonable performance on the validation set, there were signs of overfitting. This suggests that while the model learned well from the training data, it struggled to generalize unseen data, which highlights the need for additional techniques such as regularization or transfer learning to improve its performance and robustness.

CNN Model with Dropout and Regularization

To enhance generalization and reduce overfitting, adjustments were made in the next experimentation. Dropout layers and L2 regularization were introduced to prevent overfitting, which was observed in the initial model [63], [64]. Additionally, the input was reshaped to better capture patterns across all 11 images per case, ensuring that the model learned from the full dataset.

The model was refined by adding dropout layers after each convolutional layer with a dropout rate of 25%, forcing the network to rely on a broader range of features. L2 regularization was applied to the final dense layer to penalize large weights and reduce overfitting. The complexity of the model was increased by adding more convolutional layers and increasing the number of filters, enabling the model to capture finer image details.

During training, two callbacks were implemented to optimize model performance. Early stopping was used to halt training when validation loss no longer improved after 10 epochs, preventing overfitting [65]. ReduceLROnPlateau callback was employed to reduce the learning rate by a factor of 0.2 when validation loss plateaued for five epochs, promoting efficient convergence [66].

The improved model demonstrated better generalization, with a test loss of 750.287 and a MAE of 846.43. Although the model performed better in terms of overall loss, the MAE slightly increased compared to the initial model, possibly due to the added complexity. The training process ran for 23 epochs before early stopping was triggered.

Further experimentation was conducted by increasing the image dimensions to 256x256 pixels to determine if higher resolution would improve the model's predictive ability. This led to a slight change in performance, with the test loss recorded at 757.837 and the MAE at 840.66. While the model showed slight improvements in some areas, balancing complexity and performance remains a key challenge.

In the next subchapter, we will explore fine-tuning and transfer learning techniques, which offer the potential to further improve both predictive accuracy and generalization by leveraging pre-trained models on large datasets and adjusting them for our specific task.

Transfer Learning

To further improve performance, transfer learning was applied using pre-trained models, specifically MobileNetV2 and ResNet50. Both architectures have demonstrated strong performance across various medical imaging tasks, making them suitable candidates for this study [67], [68]. The primary goal was to leverage the general feature extraction capabilities of these models, which were pre-trained on the large ImageNet dataset, and adapt them for the task of predicting calcium scores from echocardiographic images [69].

To align with the input requirements of MobileNetV2 and ResNet50, the images were resized to 224x224 pixels. This step was necessary because both models were originally trained on the ImageNet dataset with images of this size. In the transfer learning process, the pre-trained layers were frozen to retain the generic features learned during initial training. Only the top layers were modified for the specific task of regression, predicting the calcium score from the echocardiographic images.

In the experiments, both models were evaluated with and without Global Average Pooling (GAP). The inclusion of GAP reduces spatial dimensions by averaging the feature maps, providing a condensed representation before the fully connected layers [70].

MobileNetV2 with GAP yielded a test loss of 1145.031 and an MAE of 1033.22. Without GAP, the test loss slightly increased to 1176.978, but the MAE improved to 926.57, suggesting that preserving spatial information led to better predictions.

For ResNet50, the model with GAP produced a test loss of 1027.619 and an MAE of 942.50. Without GAP, the test loss increased to 1780.088, though the MAE improved to 875.06. This suggests that while GAP improved the overall error metrics, the model without GAP may have been more effective at predicting the calcium score for individual cases.

In both models, transfer learning demonstrated the ability to efficiently extract useful features from the images, leveraging pre-trained knowledge while reducing training times. The results indicate that the version of ResNet50 with GAP was most effective in minimizing the overall loss, while the MobileNetV2 model without GAP performed better in terms of MAE, meaning it was more accurate on individual predictions. Further improvement can be achieved by fine-tuning the pre-trained layers, which will be explored in the next section. This would allow the models to adapt more closely to the specific patterns present in the echocardiographic data, likely enhancing their predictive accuracy.

Fine-tuning

Following the transfer learning experiments conducted with MobileNetV2 and ResNet50, we explored fine-tuning these models to further improve the performance of our calcium score prediction. Fine-tuning involves unfreezing certain layers of the pre-trained model and retraining them along with the new task-specific layers [71]. By allowing the previously frozen layers to adjust to the specific characteristics of echocardiographic images, we aimed to enhance the model's ability to capture the subtle patterns of calcification present in the dataset.

Fine-Tuning MobileNetV2

The first step was to fine-tune MobileNetV2, which had shown promising results in the transfer learning phase but still exhibited a relatively high-test MAE. In this process, we unfroze the last few convolutional layers of the pre-trained MobileNetV2 and trained them alongside the top layers specific to our calcium score prediction task. We set the learning rate to a lower value (1e-5) to avoid disrupting the pre-trained weights significantly while allowing the model to fine-tune itself to the specific features in the echocardiographic images.

The results after fine-tuning showed a marked improvement. The test loss decreased to 4293629.5, and the MAE was reduced to 1833.48, a substantial improvement from the initial transfer learning results. This suggests that fine-tuning enabled the model to better capture calcium-related features in the images, improving predictive performance.

The following plot (Figure 17) shows the training and validation loss for MobileNetV2 during fine-tuning. It can be observed that the validation loss decreases steadily, indicating that the model continues to improve in its ability to generalize to unseen data. The mean absolute error plot similarly shows a declining trend, which supports the conclusion that fine-tuning contributed positively to the model's performance.

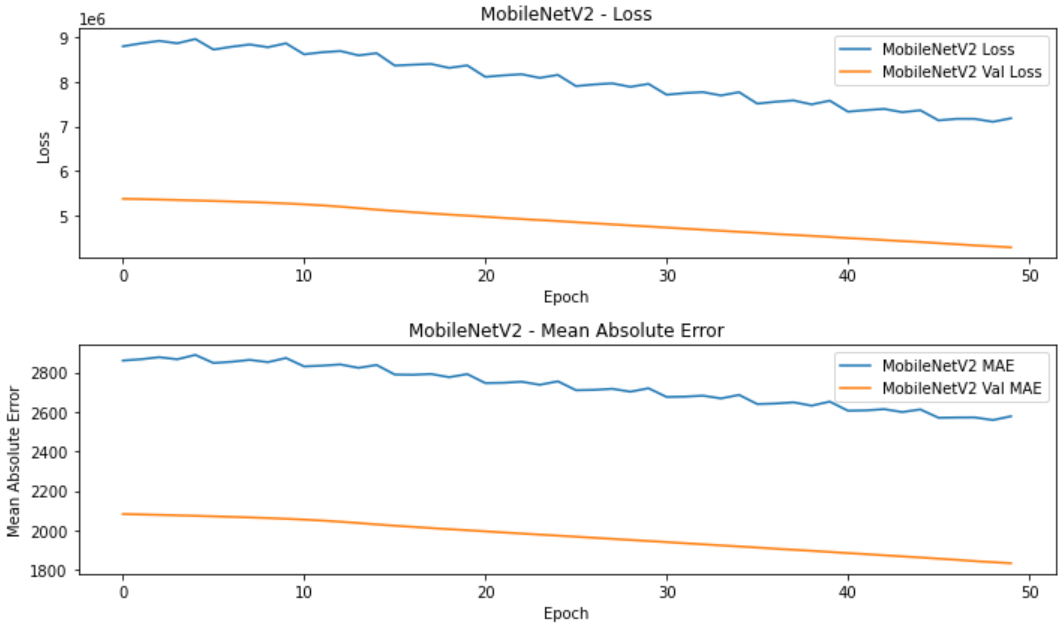

Figure 16 - MobileNetV2 Fine-Tuned Loss and MAE

Fine-Tuning ResNet50

We applied a similar fine-tuning process to ResNet50, which had outperformed MobileNetV2 in the transfer learning stage. As with MobileNetV2, we unfroze the last few layers of ResNet50 and trained them with the same lower learning rate. Given the more complex architecture of ResNet50, we expected the model to benefit even more from fine-tuning, especially considering that the echocardiographic images might contain intricate features requiring deeper feature extraction capabilities.

The fine-tuning of ResNet50 resulted in even more significant improvements. The test loss dropped to 2596564.75, and the MAE was reduced to 1356.56. This demonstrates that ResNet50, with its deeper architecture and enhanced feature extraction capabilities, benefited greatly from the fine-tuning process. The results indicate that ResNet50 is better suited for this task, particularly when fine-tuned, as it provides superior performance over MobileNetV2 in terms of both loss and error metrics.

The training and validation loss for ResNet50 (Figure 18) shows a clear downward trend, with the validation loss continuing to decrease over time. Similarly, the mean absolute error shows a notable reduction, highlighting the benefits of fine-tuning this architecture. The comparison between MobileNetV2 and ResNet50 in both plots further emphasizes ResNet50's superior performance in fine-tuning.

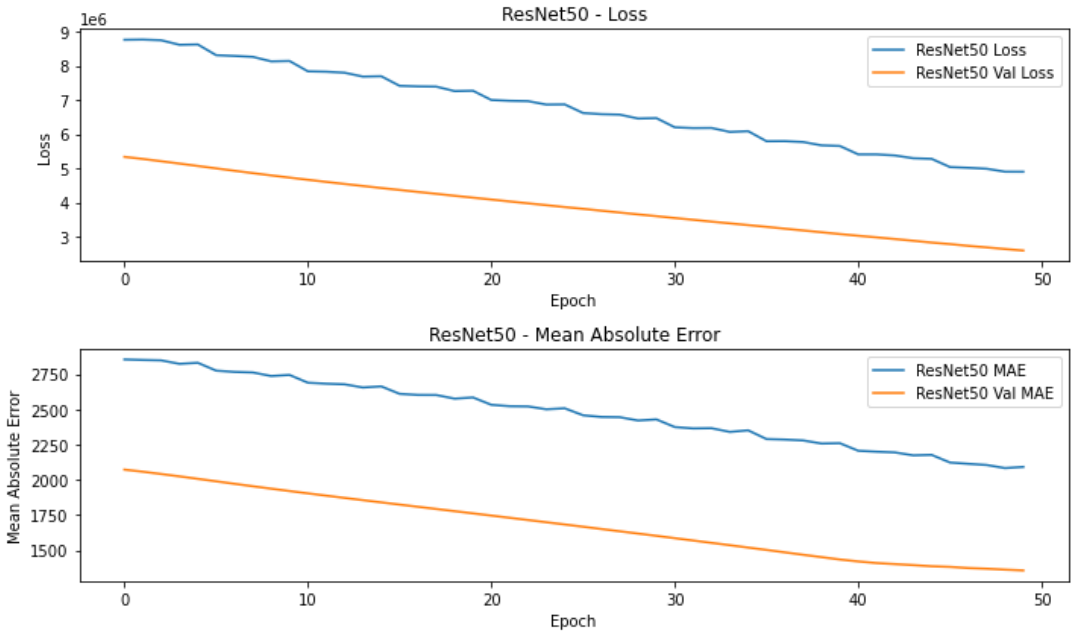

Figure 17 - ResNet50 Fine-Tuned Loss AND MAE

Both MobileNetV2 and ResNet50 showed considerable improvements after fine-tuning. However, the results clearly indicate that ResNet50 was able to extract more relevant features from the echocardiographic images, leading to better overall performance. This can be attributed to ResNet50's deeper architecture, which is better equipped to capture complex features that are likely indicative of calcification patterns in the images.

While both models demonstrated improved performance with fine-tuning, the difference in MAE suggests that deeper architectures like ResNet50 are more suitable for the task of calcium score prediction. This finding is consistent with the notion that deeper networks are more capable of capturing intricate details in medical imaging, especially in cases where subtle image features can have a significant impact on the prediction.

Fine-tuning the MobileNetV2 and ResNet50 models provided valuable improvements in calcium score prediction accuracy. While both models benefited from this process, ResNet50 emerged as the superior architecture in this context, achieving a lower test loss and MAE. The fine-tuning process allowed both models to better adapt to the nuances of echocardiographic images, leading to more accurate predictions of calcium scores.

Table 4 summarizes the performance of all the CNN models evaluated, including both initial experiments and those involving transfer learning and fine-tuning.

Table 4 - Summary of CNN Model Performance for Calcium Score Prediction

Model	Image Size	Test Loss	MAE	Comments
Initial CNN Model	128x128	1021.416	726.71	Overfitting observed
CNN with Dropout & Regularization	128x128	750.287	846.43	Improved generalization, but higher MAE
CNN with Larger Image Size	256x256	757.837	840.66	Minimal improvement despite larger image size
MobileNetV2 (with GAP)	224x224	1145.031	1033.22	Standard transfer learning performance
MobileNetV2 (without GAP)	224x224	1176.978	926.57	Better MAE but slightly worse overall loss
ResNet50 (with GAP)	224x224	1027.619	942.50	Lower test loss, higher MAE
ResNet50 (without GAP)	224x224	1780.088	875.06	Better MAE, worse overall test loss
Fine-Tuned MobileNetV2	224x224	4293629.5	1833.48	Marked improvement after fine-tuning
Fine-Tuned ResNet50	224x224	2596564.75	1356.56	Best performance, lower test loss and MAE

In comparison with the heuristic and machine learning models, the CNN approach represented a significant contribution by leveraging image data for calcium score prediction. While the initial CNN model showed signs of overfitting, the experimentation with dropout, regularization, and transfer learning ultimately led to improvements in prediction accuracy, with fine-tuned models achieving the best results. This approach offered a more direct use of echocardiographic images, setting it apart from the simpler models that relied only on pixel counts.

Patient Screening for Calcification Severity

Beyond individual model performance, the CNN models developed in this study hold potential for practical clinical application in the screening of patients based on the severity of their aortic valve calcification. By automating calcium score predictions, these models could efficiently identify patients with significant calcifications who may require further diagnostic evaluation or intervention. For instance, patients flagged by the CNN as having higher calcium scores could be prioritized for CT scans or other imaging modalities, allowing clinicians to focus resources on high-risk cases. This type of model could be particularly beneficial in streamlining the clinical workflow, enabling early detection and potentially reducing the overall burden on healthcare systems.

CHAPTER 5

Conclusion

This dissertation aimed to achieve two main objectives: detecting and extracting the image region corresponding to the aortic valve and the quantification of calcium deposits within the segmented valve. Through a combination of deep learning models, including the YOLOv8 object detection framework and CNNs, the research demonstrated the feasibility of automating the analysis of echocardiographic images for calcification detection and quantification.

In addressing RQ1, the YOLOv8n model demonstrated strong performance in detecting the aortic valve, achieving a mAP@50 of 92.88%, a precision of 99.94%, and a recall of 81.82%. These results highlight the model's effectiveness in identifying the aortic valve across diverse echocardiographic images. However, the mAP@50-95 score of 43.55% indicates that there is still potential for improvement, particularly in handling more stringent intersection-over-union (IoU) thresholds.

Regarding RQ2**,** the heuristic method developed for calcium scoring provided a fast and computationally efficient way to estimate the score by summing white pixels in binarized images that correspond to calcified regions. This method was particularly effective for quick assessments in clinical settings and demonstrated a Pearson correlation of 0.75 with calcium scores derived from CT scans. However, the method showed variability for higher calcium scores, which can be attributed to noise and limitations in accurately quantifying extensive calcifications.

A gender-based analysis further revealed that male patients tended to exhibit higher calcium scores compared to female patients. This is consistent with prior cardiovascular research, reinforcing the notion that men are more prone to severe calcific aortic valve disease.

CNN models were also applied for predicting calcium scores from multiple echocardiographic images per patient. While the multi-image input architecture provided reasonable results, overfitting was observed, leading to the introduction of regularization techniques such as dropout and L2 regularization. The refined models achieved a test loss of 750.287 and a MAE of 846.43, although additional improvements were possible with more sophisticated approaches.

The use of transfer learning with pre-trained models, such as MobileNetV2 and ResNet50, demonstrated the potential of leveraging existing models for feature extraction in medical imaging. ResNet50 performed well, achieving a test loss of 2596564.75 and an MAE of 1356.56 after fine-tuning. Fine-tuning these models allowed them to adapt more closely to

the nuances of echocardiographic data, resulting in improved calcium score predictions. This finding underscores the effectiveness of transfer learning for this type of task.

Additionally, the developed calcium quantification models have the potential to contribute to patient screening by identifying individuals with severe calcification, which is an indicator of potential aortic stenosis. By reliably flagging patients with significant calcification, this screening could assist clinicians in prioritizing those who need urgent evaluation and treatment. Early identification of patients with severe calcification could enable timely interventions, reducing the risk of complications such as aortic valve dysfunction. This approach offers a radiation-free alternative to traditional CT-based imaging, improving patient safety, streamlining clinical workflows, and enhancing the ability to monitor at-risk patients more efficiently, ultimately leading to better treatment outcomes.

The methods developed in this thesis offer a foundation for automating calcium score predictions, which could assist clinicians in making more informed decisions about treatment, especially when combined with established diagnostic tools like CT scans. Furthermore, the gender-based differences in calcium scores observed in the study underscore the importance of considering these factors in both clinical research and practice. The lower scores observed in women suggest that alternative approaches may be necessary to ensure accurate detection and risk stratification in female patients.

5.1. Limitations

While the research achieved its objectives, several limitations must be acknowledged. First, the dataset used for both training and validation was relatively small, which may have limited the ability of more complex models, such as Random Forest or Gradient Boosting, to generalize effectively. Expanding the dataset would likely lead to better performance across all models, especially for deep learning architectures.

Another limitation relates to overfitting observed in some of the initial CNN models. Although regularization techniques like dropout and L2 regularization were applied, further experimentation with alternative architectures, such as more complex network designs or ensemble methods, may be necessary to effectively mitigate overfitting and improve the model's generalization capabilities.

Additionally, the reliance on manually labeled data introduces some variability, as errors or biases in the annotations can affect model performance. While the calcium scores

derived from CT scans were used as ground truth, differences between CT and echocardiography modalities may also affect the reliability of the comparison.

5.2. Future Work

There are several promising directions for future work. Expanding the dataset to include more patients and more diverse calcification patterns will likely improve the generalization of the models. Also, incorporating patient screening based on the presence or severity of calcium deposits could enhance the clinical application of these models. For example, the models could be fine-tuned to prioritize patients with higher calcium scores, providing an automated triaging system for clinicians to identify high-risk individuals.

Moreover, exploring more complex architectures such as U-Net or attention-based mechanisms could significantly enhance segmentation accuracy, especially in cases of diffuse or subtle calcifications. These models could also be optimized to better handle variations in calcification severity, improving their precision and applicability across different patient demographics.

Experimenting with advanced CNN architectures or ensemble methods may help improve the precision of calcium score predictions. Experimenting with different image resolution settings, augmentation techniques, and deeper network layers could improve the model's ability to capture subtle calcification patterns.

To maximize the real-world clinical impact, future research should also consider integrating these models into existing diagnostic workflows. Developing user-friendly software tools or embedding the models into echocardiography imaging systems could facilitate their adoption in clinical practice and streamline the diagnostic process. Additionally, implementing patient screening as part of this process could assist clinicians in identifying which patients require closer monitoring or more frequent follow-ups based on the model's predictions of calcification severity.

5.3. Final Remarks

In conclusion, this research successfully addressed both research questions, demonstrating the potential of deep learning models, particularly YOLOv8 and CNNs, for automating the detection and quantification of calcifications in echocardiographic images. While challenges such as overfitting and limited dataset size remain, the results offer a promising foundation for future improvements in cardiovascular imaging.

Bibliography

[1] F. Kitsios, M. Kamariotou, A. I. Syngelakis, and M. A. Talias, 'Recent Advances of Artificial Intelligence in Healthcare: A Systematic Literature Review', *Applied Sciences*, vol. 13, no. 13, Art. no. 13, Jan. 2023, doi: 10.3390/app13137479.

[2] 'High-performance medicine: the convergence of human and artificial intelligence - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30617339/

[3] A. Esteva *et al.*, 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[4] 'Artificial intelligence in healthcare: past, present and future | Stroke and Vascular Neurology'. Accessed: Aug. 24, 2024. [Online]. Available: https://svn.bmj.com/content/2/4/230

[5] 'Executive Summary: Heart Disease and Stroke Statistics--2016 Update: A Report From the American Heart Association - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26811276/

[6] 'Coronary artery calcium and primary prevention risk assessment: what is the evidence? An updated meta-analysis on patient and physician behavior - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22811506/

[7] 'Long-term prognosis associated with coronary calcification: observations from a registry of 25,253 patients - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17481445/

[8] 'Coronary calcium as a predictor of coronary events in four racial or ethnic groups - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18367736/

[9] '2017 ESC/EACTS Guidelines for the management of valvular heart disease - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28886619/

[10] 'Appropriate Use Criteria for Echocardiography. A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21349406/

[11] 'Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25559473/

[12] 'Automated Echocardiographic Detection of Severe Coronary Artery Disease Using Artificial Intelligence - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34922865/

[13] 'Fully Automated Echocardiogram Interpretation in Clinical Practice - PubMed'. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30354459/

[14] F. Yang *et al.*, 'Automated Analysis of Doppler Echocardiographic Videos as a Screening Tool for Valvular Heart Diseases', *JACC: Cardiovascular Imaging*, vol. 15, no. 4, pp. 551–563, 2022, doi: 10.1016/j.jcmg.2021.08.015.

[15] 'Aortic Valve Segmentation using Deep Learning | IEEE Conference Publication | IEEE Xplore'. Accessed: Aug. 27, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9398738

[16] 'Multi-center retrospective cohort study applying deep learning to electrocardiograms to identify left heart valvular dysfunction'. Accessed: Sep. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/368508632_Multi-center_retrospective_cohort_study_applying_deep_learning_to_electrocardiograms_to_identify_left_heart_valvular_dysfunction

[17] 'Self-supervised contrastive learning of echocardiogram videos enables label-efficient cardiac disease diagnosis'. Accessed: Sep. 05, 2024. [Online]. Available: https://arxiv.org/abs/2207.11581

[18] 'A deep learning algorithm accurately detects pericardial effusion on echocardiography | Request PDF'. Accessed: Sep. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/339964837_a_deep_learning_algorithm_accurately_detects_ pericardial_effusion_on_ echocardiography

[19] P. Gać *et al.*, ‘Aortic Valve Calcium Score: Applications in Clinical Practice and Scientific Research—A Narrative Review’, *JCM*, vol. 13, no. 14, p. 4064, Jul. 2024, doi: 10.3390/jcm13144064.
[20] ‘(PDF) A Systematic Literature Review on Applying CRISP-DM Process Model’. Accessed: Aug. 24, 2024. [Online]. Available: https://www.researchgate.net/publication/349527794_A_Systematic_Literature_Review_on_Applying_CRISP-DM_Process_Model
[21] ‘The PRISMA 2020 statement: an updated guideline for reporting systematic reviews - PubMed’. Accessed: Aug. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33782057/
[22] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, ‘Rayyan—a web and mobile app for systematic reviews’, *Syst Rev*, vol. 5, no. 1, p. 210, Dec. 2016, doi: 10.1186/s13643-016-0384-4.
[23] M. Nizar, C. Chan, A. Khalil, A. Yusof, and K. Lai, ‘Real-time Detection of Aortic Valve in Echocardiography using Convolutional Neural Networks’, *CURRENT MEDICAL IMAGING*, vol. 16, no. 5, pp. 584–591, 2020, doi: 10.2174/1573405615666190114151255.
[24] W. Dai, H. Nazzari, M. Namasivayam, J. Hung, and C. M. Stultz, ‘Identifying Aortic Stenosis With a Single Parasternal Long-Axis Video Using Deep Learning’, *Journal of the American Society of Echocardiography*, vol. 36, no. 1, pp. 116–118, 2023, doi: 10.1016/j.echo.2022.10.014.
[25] G. Holste *et al.*, ‘Severe aortic stenosis detection by deep learning applied to echocardiography’, *European Heart Journal*, vol. 44, no. 43, pp. 4592–4604, 2023, doi: 10.1093/eurheartj/ehad456.
[26] S. V. Wifstad *et al.*, ‘Quantifying Valve Regurgitation Using 3-D Doppler Ultrasound Images and Deep Learning’, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 12, pp. 3317–3326, 2022, doi: 10.1109/TUFFC.2022.3218281.
[27] L. B. Elvas, A. G. Almeida, L. Rosario, M. S. Dias, and J. C. Ferreira, ‘Calcium identification and scoring based on echocardiography. An exploratory study on aortic valve stenosis’, *Journal of Personalized Medicine*, vol. 11, no. 7, 2021, doi: 10.3390/jpm11070598.
[28] L. Tang, X. Wang, J. Yang, Y. Wang, M. Qu, and H. Li, ‘DLFFNet: A new dynamical local feature fusion network for automatic aortic valve calcification recognition using echocardiography’, *Computer Methods and Programs in Biomedicine*, vol. 243, 2024, doi: 10.1016/j.cmpb.2023.107882.
[29] S. Kim *et al.*, ‘Assessment of valve regurgitation severity via contrastive learning and multi-view video integration’, *Physics in Medicine and Biology*, vol. 69, no. 4, 2024, doi: 10.1088/1361-6560/ad22a4.
[30] K. R. Steffner *et al.*, ‘Deep learning for transesophageal echocardiography view classification’, *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-023-50735-8.
[31] N. Yuan *et al.*, ‘Prediction of Coronary Artery Calcium Using Deep Learning of Echocardiograms’, *Journal of the American Society of Echocardiography*, vol. 36, no. 5, pp. 474-481.e3, 2023, doi: 10.1016/j.echo.2022.12.014.
[32] A. Karužas *et al.*, ‘Artificial intelligence for automated evaluation of aortic measurements in 2D echocardiography: Feasibility, accuracy, and reproducibility’, *Echocardiography*, vol. 39, no. 11, pp. 1439–1445, 2022, doi: 10.1111/echo.15475.
[33] R. Nath, A. Kazemi, S. Callahan, M. F. Stoddard, and A. A. Amini, ‘4Dflow-VP-Net: A deep convolutional neural network for noninvasive estimation of relative pressures in stenotic flows from 4D flow MRI’, *Magnetic Resonance in Medicine*, vol. 90, no. 5, pp. 2175–2189, 2023, doi: 10.1002/mrm.29791.
[34] N. Ahmadi, M. Y. Tsang, A. N. Gu, T. S. M. Tsang, and P. Abolmaesumi, ‘Transformer-Based Spatio-Temporal Analysis for Classification of Aortic Stenosis Severity from Echocardiography Cine Series’, *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 366–376, 2024, doi: 10.1109/TMI.2023.3305384.
[35] Y. Lei *et al.*, ‘Echocardiographic image multi-structure segmentation using Cardiac-SegNet’, *Medical Physics*, vol. 48, no. 5, pp. 2426–2437, 2021, doi: 10.1002/mp.14818.
[36] Z. Liu *et al.*, ‘Automated deep neural network-based identification, localization, and tracking of cardiac structures for ultrasound-guided interventional surgery’, *Journal of Thoracic Disease*, vol. 15, no. 4, pp. 2129–2140, 2023, doi: 10.21037/jtd-23-470.

[37] S. A. Coulter and K. Campos, ‘Artificial Intelligence in Echocardiography’, *Texas Heart Institute Journal*, vol. 49, no. 2, 2022, doi: 10.14503/THIJ-21-7671.
[38] O. S. Pianykh, *Digital imaging and communications in medicine (DICOM): A practical introduction and survival guide*. 2008, p. 383. doi: 10.1007/978-3-540-74571-6.
[39] ‘Real-Time Computer Vision with OpenCV | Request PDF’. Accessed: Sep. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/262368549_Real-Time_Computer_Vision_with_OpenCV
[40] ‘Array programming with NumPy - PubMed’. Accessed: Sep. 05, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32939066/
[41] ‘Matplotlib: A 2D Graphics Environment’. Accessed: Sep. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/3422921_Matplotlib_A_2D_Graphics_Environment
[42] ‘Journal of Open Source Software: DICaugment: A Python Package for 3D Medical Imaging Augmentation’. Accessed: Sep. 05, 2024. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.06120
[43] C. Shorten and T. M. Khoshgoftaar, ‘A survey on Image Data Augmentation for Deep Learning’, *J Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[44] L. Perez and J. Wang, ‘The Effectiveness of Data Augmentation in Image Classification using Deep Learning’, 2017, *arXiv*. doi: 10.48550/ARXIV.1712.04621.
[45] S. Gomes, L. B. Elvas, J. C. Ferreira, and T. Brandão, ‘Automatic Calcium Detection in Echocardiography Based on Deep Learning: A Systematic Review’, in *Innovations in Bio-Inspired Computing and Applications*, vol. 649, A. Abraham, A. Bajaj, N. Gandhi, A. M. Madureira, and C. Kahraman, Eds., in Lecture Notes in Networks and Systems, vol. 649. , Cham: Springer Nature Switzerland, 2023, pp. 754–764. doi: 10.1007/978-3-031-27499-2_70.
[46] ‘Jocher, G., Chaurasia, A. and Qiu, J. (2023) YOLO by Ultralytics. - References - Scientific Research Publishing’. Accessed: Sep. 03, 2024. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=3532980
[47] ‘A survey on deep learning in medical image analysis - ScienceDirect’. Accessed: Sep. 03, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1361841517301135
[48] ‘Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? - PubMed’. Accessed: Sep. 03, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29994302/
[49] S. K. Zhou *et al.*, ‘A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises’, *Proc IEEE Inst Electr Electron Eng*, vol. 109, no. 5, pp. 820–838, May 2021, doi: 10.1109/JPROC.2021.3054390.
[50] ‘Statistical Power Analysis for the Behavioral Sciences | ScienceDirect’. Accessed: Sep. 07, 2024. [Online]. Available: https://www.sciencedirect.com/book/9780121790608/statistical-power-analysis-for-the-behavioral-sciences
[51] J. Zar, ‘Spearman Rank Correlation’, in *Encycl Biostat*, vol. 5, 2005. doi: 10.1002/0470011815.b2a15150.
[52] ‘Nonparametric Statistical Inference, Fifth Edition-CRC Press (2010)-Gibbons.pdf | Nicko V. - Academia.edu’. Accessed: Sep. 07, 2024. [Online]. Available: https://www.academia.edu/28630519/Nonparametric_Statistical_Inference_Fifth_Edition_CRC_Press_2010_Gibbons_pdf
[53] ‘Gender Differences in Cardiac Hypertrophy | Journal of Cardiovascular Translational Research’. Accessed: Sep. 10, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12265-019-09907-z
[54] ‘Coronary Artery Calcification; report from the Multi-Ethnic Study of Atherosclerosis - PMC’. Accessed: Sep. 10, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5108577/
[55] O. Manfrini *et al.*, ‘Sex Differences in Modifiable Risk Factors and Severity of Coronary Artery Disease’, *JAHA*, vol. 9, no. 19, p. e017235, Oct. 2020, doi: 10.1161/JAHA.120.017235.
[56] J. J. Thaden *et al.*, ‘Sex-related differences in calcific aortic stenosis: correlating clinical and echocardiographic characteristics and computed tomography aortic valve calcium score to excised aortic valve weight’, *Eur Heart J*, vol. 37, no. 8, pp. 693–699, Feb. 2016, doi: 10.1093/eurheartj/ehv560.

[57] A. A. Kelkar *et al.*, ‘Long-Term Prognosis After Coronary Artery Calcium Scoring Among Low-Intermediate Risk Women and Men’, *Circ: Cardiovascular Imaging*, vol. 9, no. 4, p. e003742, Apr. 2016, doi: 10.1161/CIRCIMAGING.115.003742.
[58] H. Zou and T. Hastie, ‘Regularization and Variable Selection Via the Elastic Net’, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.
[59] ‘The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) | Request PDF’. Accessed: Sep. 10, 2024. [Online]. Available: https://www.researchgate.net/publication/319770542_The_Elements_of_Statistical_Learning_Data_Mining_Inference_and_Prediction_Second_Edition_Springer_Series_in_Statistics
[60] T. Chai and R. R. Draxler, ‘Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature’, *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.
[61] J. Miles, ‘R Squared, Adjusted R Squared’, in *Wiley StatsRef: Statistics Reference Online*, 1st ed., R. S. Kenett, N. T. Longford, W. W. Piegorsch, and F. Ruggeri, Eds., Wiley, 2014. doi: 10.1002/9781118445112.stat06627.
[62] D. P. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimization’, 2014, *arXiv*. doi: 10.48550/ARXIV.1412.6980.
[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ‘Dropout: A Simple Way to Prevent Neural Networks from Overfitting’, *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
[64] A. Y. Ng, ‘Feature selection, L_1 vs. L_2 regularization, and rotational invariance’, in *Twenty-first international conference on Machine learning - ICML ’04*, Banff, Alberta, Canada: ACM Press, 2004, p. 78. doi: 10.1145/1015330.1015435.
[65] L. Prechelt, ‘Early Stopping - But When?’, in *Neural Networks: Tricks of the Trade*, vol. 1524, G. B. Orr and K.-R. Müller, Eds., in Lecture Notes in Computer Science, vol. 1524. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 55–69. doi: 10.1007/3-540-49430-8_3.
[66] L. N. Smith, ‘Cyclical Learning Rates for Training Neural Networks’, in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA: IEEE, Mar. 2017, pp. 464–472. doi: 10.1109/WACV.2017.58.
[67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, ‘MobileNetV2: Inverted Residuals and Linear Bottlenecks’, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
[68] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep Residual Learning for Image Recognition’, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[69] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, ‘ImageNet: A large-scale hierarchical image database’, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[70] M. Lin, Q. Chen, and S. Yan, ‘Network In Network’, 2013, *arXiv*. doi: 10.48550/ARXIV.1312.4400.
[71] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, ‘How transferable are features in deep neural networks?’, 2014, doi: 10.48550/ARXIV.1411.1792.