INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Automatic Calcium Detection in Echocardiography Based on Deep Learning

Sara Raquel de Sá Gomes

Master's in Integrated Business Intelligence Systems

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

Co-supervisor:
MSc Luís Manuel Nobre de Brito Elvas, Invited Assistant,
ISCTE-IUL

October, 2023

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Automatic Calcium Detection in Echocardiography Based on Deep Learning

Sara Raquel de Sá Gomes

Master's in Integrated Business Intelligence Systems

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

Co-supervisor:
MSc Luís Manuel Nobre de Brito Elvas, Invited Assistant,
ISCTE-IUL

October, 2023

“Remember to look up at the stars and not down at your feet.”

Stephen Hawking

Acknowledgements

I want to thank all the people who contributed for the development of this dissertation, not only with direct technical guidance, but also for supporting and encouraging me to reach this objective.

To my supervisors, Tomás Brandão and Luís Elvas, for all the knowledge sharing and with whom I learned so much in the last 2 years. Their availability and wise advice were fundamental to guide me in this journey that I am proud of finished. I am extremely thankful to them.

To professor João Carlos Ferreira, this master's coordinator, who has accompanied my progress since the beginning, not only in this dissertation but also along the entire master's course. Always giving great advice and support to the students.

To professor Luís Rosário, from Santa Maria's Hospital, who demonstrated a big interest in this dissertation development, always having a positive word to say, even when facing an obstacle. His collaboration and knowledge were fundamental since the data acquisition phase through the results evaluation.

To my colleagues for all the experience sharing through this master's course, especially to my friend André Bergano, for the mutual support, for the honest opinions and for all the concerns releasing.

To my parents and sister, for helping me and giving me all the emotional support I needed. For raising me with so much love and for all the dedication during my life. Without them, I would not be here.

To my boyfriend João Pedro, who always believed in me and encouraged me to not give up. For appearing in my life and turn it better, for the patience during this step and for making me feel strong.

To conclude, I want to thank all my family and friends, and dedicate this dissertation to all these people that are part of my life and contributed in some way to conclude this step. I am very thankful to all of them.

Resumo

As doenças cardiovasculares são uma das principais causas de morte em todo o mundo. Exames como ecocardiogramas ou Tomografias Computorizadas (TACs), realizados por um médico especialista, é um meio de diagnóstico essencial para detetar várias doenças cardíacas. Podem ser utilizados modelos de Aprendizagem Profunda para desenvolver ferramentas orientadas capazes de processar e categorizar este tipo de dados, fornecendo diagnósticos rápidos com elevada confiança, que podem tornar o trabalho dos médicos mais eficiente e melhorar o sistema de cuidados de saúde. Foi realizada uma revisão sistemática das aplicações de Aprendizagem Profunda em imagens médicas, que concluiu que existem alguns desenvolvimentos nesta área, principalmente recorrendo a TACs, embora este tipo de exames apresente algumas desvantagens, como a exposição a radiação de risco. A ecocardiografia, apesar de ainda pouco explorada, pode ser uma excelente alternativa. Nesta dissertação foi desenvolvido um método automático de deteção de Estenose Aórtica em Ecocardiografias, baseado em Redes Neuronais Convolucionais (CNNs), constituído essencialmente por dois processos: (1) um detetor de objetos para localizar a válvula aórtica – que atingiu 95% de precisão e 100% de revocação; e (2) um classificador para identificar depósitos de cálcio na válvula – cujos resultados atingiram 92% de precisão e 100% de revocação, sendo também muito promissores. Este desenvolvimento representa uma melhoria na eficiência do fluxo de trabalho dos sistemas de saúde, conduzindo a diagnósticos mais rápidos e fiáveis e, consequentemente, a uma diminuição da mortalidade por fatores cardíacos.

Abstract

Cardiovascular diseases are one of the main causes of death in the world. The examination of echocardiography or Computed Tomography (CT) images, which a health care specialist primarily does, is an important way to diagnose several cardiac diseases. Deep Learning (DL) algorithms can be used to develop a data-driven tool that can process and categorize this kind of data, providing fast diagnoses with considered reliability, to aid physicians in their work and increase healthcare effectiveness. A systematic review of DL applications on health images concluded that there are some developments in this field primarily using CT scans, although they have some disadvantages like radiation exposure risk. Echocardiography may be a great alternative, not explored yet. In this dissertation a fully automated method based on Convolutional Neural Networks (CNNs) to detect Aortic Stenosis (AS) in echocardiography images was developed, consisting in two essential processes: (1) an object detector to locate aortic valve – achieving 95% precision and 100% recall; and (2) a classifier to identify calcium structures in the valve – which achieved promising results of 92% precision and 100% recall. The outcome of this work is an improvement in the efficiency of health systems' workflow, leading to faster and reliable diagnoses and, consequently, a decrease in deaths from heart diseases.

Table of contents

Chapter 1 1

1.1. Context 1

1.2. Motivation 2

1.3. Methodology 3

1.4. Objectives 4

Chapter 2 5

2.1. Systematic Literature Procedures 5

2.2. Retrieved Documents 6

2.3. Related Work 9

2.4. Summary 13

Chapter 3 15

3.1. Data Understanding 15

3.2. Pre-processing 15

3.3. Annotations 18

3.4. Data Augmentation 19

Chapter 4 21

4.1. Aortic Valve Detection 22

4.1.1. Object Detector Modeling 22

4.1.2. Performance Metrics 23

4.1.3. Experimental Procedures 25

4.2. Aortic Stenosis Classification 32

4.2.1. Classifier Modeling 32

4.2.2. Performance Metrics 33

4.2.3. Experimental Procedures 33

4.3. Discussion 40

Chapter 5 43

5.1. Final Remarks 43

5.2. Future Work 44

List of Figures, Tables and Equations

Figure 1 - CRISP-DM methodology steps 3
Figure 2 - PRISMA selection process 6
Figure 3 - Keywords analysis 7
Figure 4 - ML techniques used on analyzed papers 8
Figure 5 - Imaging data types used on analyzed papers 8
Figure 6 - Application goals of the analyzed papers 9
Figure 7 - Heart detection (1) and calcium segmentation (2) in CT image 10
Figure 8 - Noise reduction in a CT-scan. The original image (A) is transformed in a noise-reducted image (B) 12
Figure 9 - Different types of heart view on Echocardiography: (A) Apical 2-chamber view; (B) Apical 3-chamber view; (C) Apical 4-chamber view; (D) Apical 5-chamber view; (E) Parasternal short-axis view; (F) Parasternal short-axis view of the left ventricle; (G) Parasternal long-axis view 17
Figure 10 - Image transformation process 18
Figure 11 - Image annotation process 19
Figure 12 - Data augmentation process. (A) Translation; (B) Zoom; (C) Rotation; (D) Gama contrast 20
Figure 13 - Framework for aortic valve detection and Aortic Stenosis classification 21
Figure 14 - Object detector with Transfer Learning for aortic valve detection 23
Figure 15 - IoU calculation 24
Figure 16 - Test set of experiment 9, where the model could not detect aortic valve 26
Figure 17 - Confusion matrix for object detection with original dataset 28
Figure 18 - Loss and metrics curves for object detection with original dataset 28
Figure 19 - Aortic valve detection with the original dataset - examples 29
Figure 20 - Confusion matrix for object detection with data augmentation (rotations) 30
Figure 21 - Loss and metrics curves for object detection with data augmentation (rotations) 30
Figure 22 - Aortic valve detection with the best model - examples 31
Figure 23 - Classification model with Transfer Learning for Aortic Stenosis classification 32
Figure 24 - Heuristic classification process 33
Figure 25 - Neural Network architecture 35
Figure 26 – Process of cropping the area of interest 35
Figure 27 - Process of noise reduction 36
Figure 29 - Accuracy and loss curves for image classification with EfficientNetB0 39
Figure 28 - Confusion matrix for image classification with EfficientNetB0 39
Figure 29 - Accuracy and loss curves for image classification with EfficientNetB0 39

Table 1 - PRISMA research results 5
Table 2 - Topics analysis from the systematic review 10
Table 3 - Aortic valve detection results using Dataset 1 25
Table 4 - Object detection results using Dataset 2 27
Table 5 - Aortic Stenosis classification results for experiments using DL .**Error! Bookmark not defined.**
Table 6 - Aortic Stenosis classification results for heuristic experiments .**Error! Bookmark not defined.**

Equation 1 - IoU calculation formula..... 23
Equation 2 - Precision calculation formula..... 24
Equation 3 - Recall calculation formula..... 24
Equation 4 - F1-score calculation formula 25
Equation 5 - Accuracy calculation formula..... 33

Glossary

AI – Artificial Intelligence

CNN – Convolutional Neural Network

CRISP-DM – Cross Industry Standard Process for Data Mining

CT – Computed Tomography

DL – Deep Learning

NN – Neural Network

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

CHAPTER 1

Introduction

In an era dominated by Artificial Intelligence (AI), questions about its practical benefits and societal implications have come to the forefront [1], [2]. This study aims to focus on the application of AI in healthcare, specifically in the field of cardiology, highlighting its possible advantages on improving diagnoses and medical workflow.

The research presented in this Dissertation is centered on two key tasks: the development of AI algorithms for aortic valve detection and the classification of valve calcification. By doing so, the aim is to shed light on how AI can enhance medical diagnostics and reduce the practitioners' workload.

1.1. Context

In 2019, cardiovascular disorders accounted for one-third of all fatalities worldwide [3], significantly affecting disability increase. Between 1990 and 2019, the number of people with cardiovascular problems doubled, from 271 million to 523 million. Additionally, deaths from cardiovascular disorders increased from 12.1 million to 18.6 million during this period [4].

According to [5], Coronary Artery Calcium (CAC) is a significant predictor of cardiovascular failure events and mortality due to its high correlation with coronary artery disorders. The formation of calcific structures inside the aortic valve leaflets is commonly thought to be the cause of Aortic Valve Sclerosis (AVS) or Aortic Valve Calcification (AVC) [6]. The primarily heart valve disease is Aortic Stenosis (AS), which is one of the major causes of cardiovascular death [7].

One of the ways to prevent cardiovascular calcifications is regular surveillance. Medical imaging is a potent diagnostic technique that provides knowledge about anatomical structures using imaging modalities including Computed Tomography (CT)-scan, Magnetic Resonance Imaging (MRI), Angiogram, Electrocardiogram (ECG), and others. Thanks to developments in biomedical imaging, digital medical image processing can contribute to the earlier diagnosis of anomalies or disorders in any system of the human body. Given that cardiology is one of the most complex areas within medicine, there are substantial technological advancements yet to be made [8].

Regular CT scans and echocardiograms are two main types of medical imaging which can be used to identify cardiovascular calcifications. Despite providing a great predictive capability because of its high-quality image, ionizing radiation exposure is one of the crucial issues with using CT scans, since it can afford long time consequences for the patients' heath [9]. It is important to balance whether the outcome from this method is an advantage contrasting to the concern about risk exposure, which leads to avoiding the use of CT scans.

On the other hand, echocardiography is becoming the first-line method for diagnosing and monitoring the treatment of AS, according to the European Society of Cardiology [10]. It is easy portability, low cost and, especially, zero radiation exposure are great advantages [10]. The pressure gradient of cardiac lesions can be easily identified in an echocardiography, and as it uses sound waves rather than radiation, it is considered a safe method [8].

1.2. Motivation

AI may positively impact each step of cardiovascular imaging, from acquisition through reporting. Automated acquisition, segmentation, classification, and report generation can become faster and more accurate with AI [11]. Coronary calcium quantification, medical imaging measurements, or detection of cardiac illnesses are a few examples where AI can be applied to provide an accurate diagnosis and reduce observer variation [12].

Convolutional Neural Networks (CNNs) are frequently used on Deep Learning (DL) models. Due to their capacity to automatically identify the most advantageous features for the application context, this type of neural network is appropriate for a variety of Machine Learning (ML)–based Computer Vision (CV) applications. However, they require vast datasets, as their performance often improves with a substantial amount of data [13]. CNNs have been employed successfully for medical image segmentation and classification [5], [14]–[18].

Numerous studies, as referenced in [19], have focused on automating calcium measurement on CT scans. Nevertheless, there remains substantial unexplored territory in the field of echocardiography, as mentioned in [20]. In this study, a semi-automatic model for calcium detection on echocardiographic images has been developed, demonstrating promising results. Consequently, it can be inferred that a fully automatic model could potentially offer a more efficient solution, avoiding the need for user training. The anticipated impact of these advancements in the field of echocardiography is to assist medical workflows by providing fast and precise diagnoses, all while minimizing the radiation exposure associated with CT scans.

1.3. Methodology

The structure of this Dissertation follows Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, the most used methodology for data mining and analytics works [21], which consists in 6 steps as shown on Figure 1 [22]:

- **Business Understanding**: Assess the business situation to understand available and needed resources; clearly define the data mining goal and success criteria; develop a mandatory project plan.
- **Data Understanding**: Collect and explore data from various sources; check data quality; use statistical analysis to describe data, identifying attributes and their relationships.
- **Data Preparation**: Define inclusion and exclusion criteria for data selection; address poor data quality through data cleaning; construct derived attributes based on the chosen model from the first phase.
- **Modeling**: Select a modeling technique based on the business problem and data; set specific parameters for building the model; evaluate the model against predefined criteria to choose the best one.
- **Evaluation**: Check results against defined business objectives; discuss results and define future work; review the overall process.
- **Deployment**: Plan the deployment phase, which could be in the form of a final report or a software component; the deployment phase involves planning, monitoring, and maintenance.

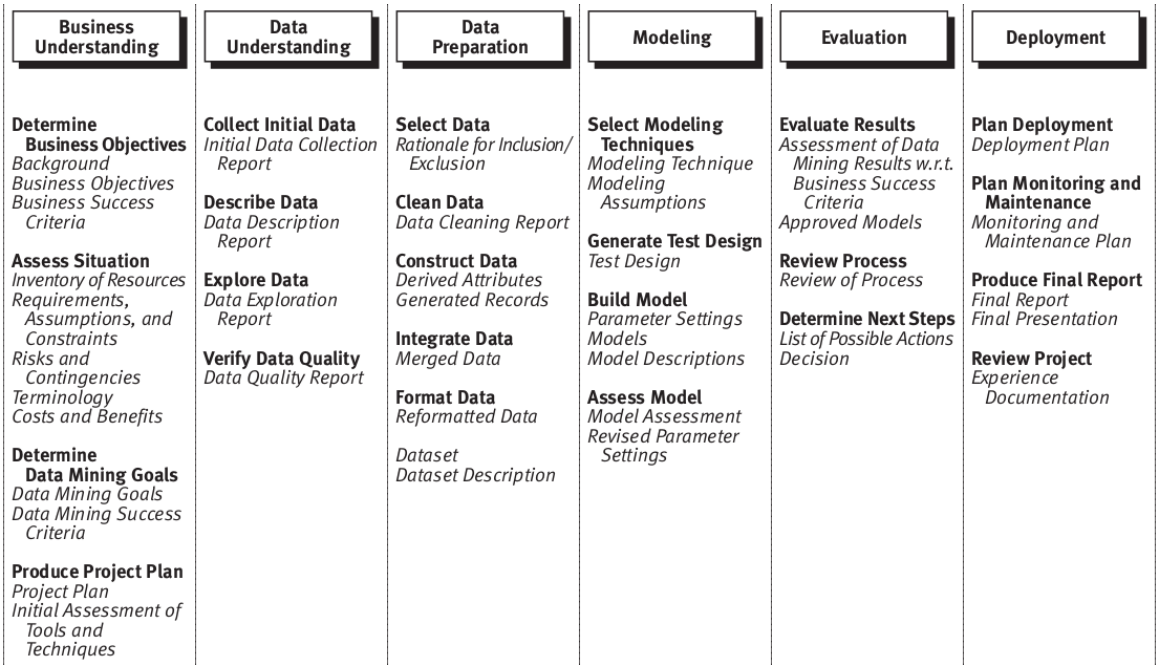


Figure 1 - CRISP-DM methodology steps

1.4. Objectives

The primary aim of this dissertation is to automate the detection of AS in patients by identifying calcium deposits in the aortic valve using echocardiographic data. To achieve this objective, this work was divided into two distinct problems: (1) the development of an object detection algorithm for identifying the aortic valve and (2) the creation of a classification algorithm to determine the presence of calcium within the valve. DL techniques were tried, using CNNs, and a manual classification approach was also followed, for results comparing.

This dissertation will be presented in five chapters (including Introduction):

- **Chapter 2:** State-of-the-art analysis, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [24], to understand the existent developments in the area.
- **Chapter 3:** Selection of data, essential to meet the objectives in modeling phase, in collaboration with Affidea clinics. Description of data preparation process, through anonymization and uniformization.
- **Chapter 4:** Building and application of different modeling techniques, trying to find the best approach. Description of all the approaches followed and respective results, divided into two sections to meet each of the objectives mentioned. Discussion and conclusions drawing about the results obtained.
- **Chapter 6:** A reflection about the work developed and improvement opportunities, and definition of possible future developments.

CHAPTER 2

State-of-the-art

2.1. Systematic Literature Procedures

To better understand what the State of the Art on DL-based CV applications to health care is, and to define techniques and methods to use on this Dissertation, a systematic literature review was performed following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, which is a guideline to assist writers in better systematic review evaluation and reporting, so that readers can comprehend the authors' research and conclusions [24]. Following PRISMA methodology, the goal of the systematic review presented in this chapter is to provide an answer to the question "What is the state-of-the-art on DL-based models for automatic calcium detection on cardiac images?".

In October 2022, research on Scopus and Web of Science Core Collection (WoSCC) databases was performed. Three categories have been established for defining searching keywords:

- Concept → Definition of techniques to be used to achieve the goal;
- Image Modality → Types of data to be used and analyzed. In this case, the exam types;
- Context → The objectives to achieve with the proposed work.

As seen in Table 1, the research expression was constructed by combining all the columns, i.e., Concept AND Image Modality AND Context. Only reviews and articles were considered, and publication years were filtered within the last four years (2019-2022).

Table 1 - PRISMA research results

<table>
<tr><th>Concept</th><th>Image Modality</th><th>Context</th><th>Limitations</th></tr>
<tr><td rowspan="2">"deep learning"
"neural network"
"computer vision"

1.004.076 + 414.288</td><td rowspan="2">"computed tomography"
"ct-scan"
"echocardiography"
"ultrassound image"</td><td rowspan="3">"aortic stenosis"
"aortic calcification"
"calcium score"
"aortic sclerosis"</td><td rowspan="4">year: 2019-2022

document type:
article or review</td></tr>
<tr></tr>
<tr><td colspan="2">11.268 + 4.078</td></tr>
<tr><td colspan="3">109 + 20</td></tr>
<tr><td colspan="4">80 + 12</td></tr>
</table>

2.2. Retrieved Documents

The research described in the previous section resulted in 92 results, 80 from Scopus and 12 from WoSCC. These 92 documents were analyzed through PRISMA filtering processes, as illustrated in Figure 2.

After dropping out duplicates, 82 documents were analyzed and labeled, to discard the ones that were out of scope. Of 50 eligible papers, 38 met the present research goals, so they were selected to be included in the literature review.

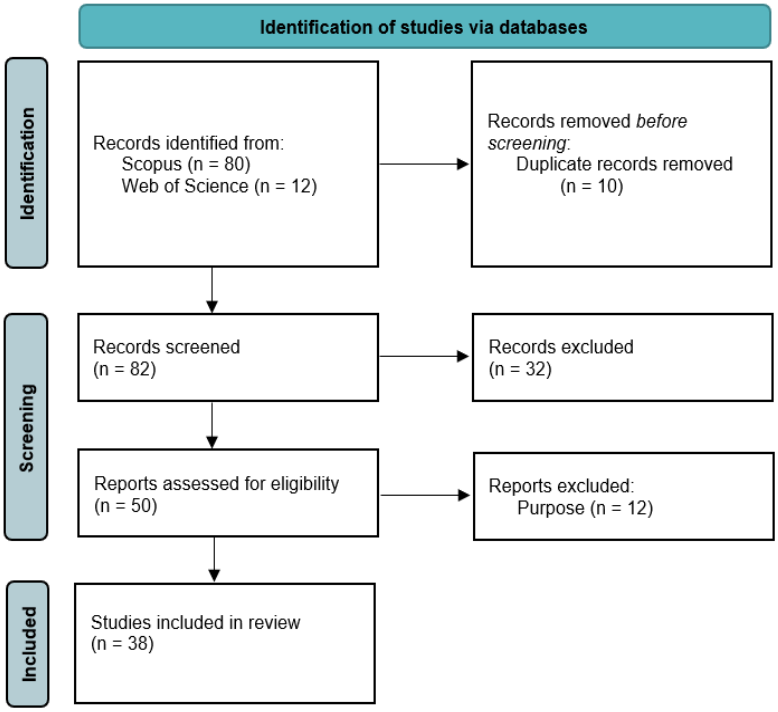


Figure 2 - PRISMA selection process

The diagram presented on Figure 3 was built using VOSviewer, and its objective is to facilitate identification of the keywords addressed to the selected documents, and correlations between them. The keywords are shown in this graphic with varying sizes according to the frequency they appear. As shown, "Deep Learning" and "Artificial Intelligence" are the two primary keywords, followed by "Cardiovascular Disease" and "Machine Learning." Colors represent clusters – groups of keywords with a strong relation between them.

Clearly, "Deep Learning" and "Computed Tomography" are related. The number of developed DL-based studies applied to echocardiography, in contrast, is still limited – there are no stronger relationships including Echocardiography keyword.

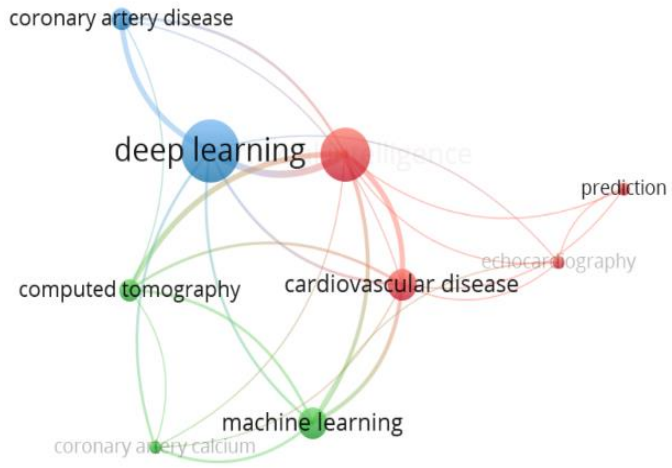


Figure 3 - Keywords analysis

Then, an analysis of the ML techniques applied, image types used, and application goal presented by each paper was performed, trying to identify developments already existent, and opportunities to improve in the area.

In the field of ML techniques, as shown in Figure 4, three types of Neural Networks (NNs) – Convolutional NNs (CNNs), Fully connected NNs (FCNNs) and Recurrent NNs (RNNs), with a clear highlight to CNNs, appear in 31 of the 38 analyzed papers (over 80%). 10 papers also addressed other ML approaches, using simpler or more traditional ML techniques instead of NNs. It should be noted that the same paper can address more than one ML technique.

As for the imaging data types used, represented in Figure 5, CT-scans were covered in 31 of the 38 analyzed documents (over 80%). There are 5 papers mentioning various types of images, such as CT-scans, MRI or Echocardiography, making a comparison between them. Only 2 papers present developments in Echocardiography, which is the type of image selected to use in this dissertation.

The application goals of the selected set of references are represented in Figure 6. The purposes are mainly Calcium quantification or segmentation and detection of diseases – AS, Covid-19 and others. Calcium quantification or segmentation is highlighted by being one of the goals of 31 papers. 12 studies had another type of objective. It should be noted that the same paper can present more than one application goal.

Looking at the intersection of the 3 groups of topics, it is observable that 22 out of the 38 papers (58%) mention CNNs, CT-scans and Calcium quantification or segmentation simultaneously. It is worth mentioning that Echocardiography was not addressed on any of the works about calcium quantification.

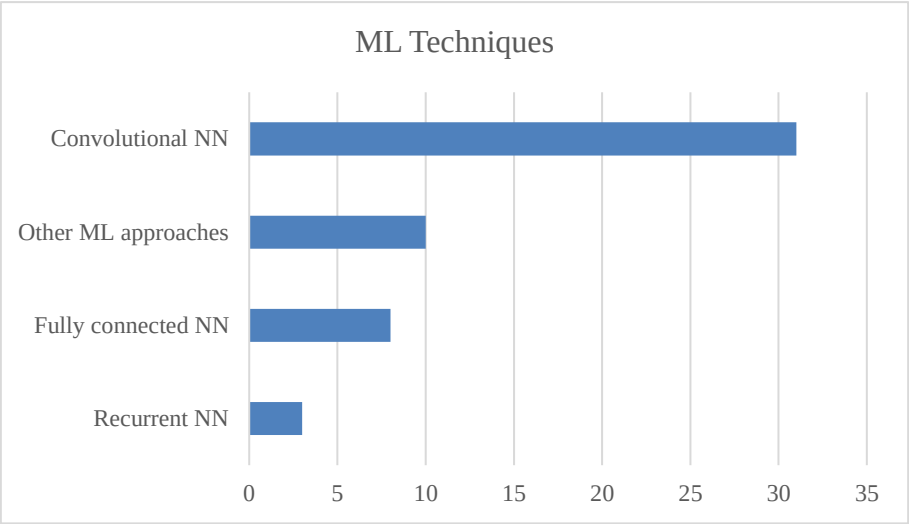


Figure 4 - ML techniques used on analyzed papers

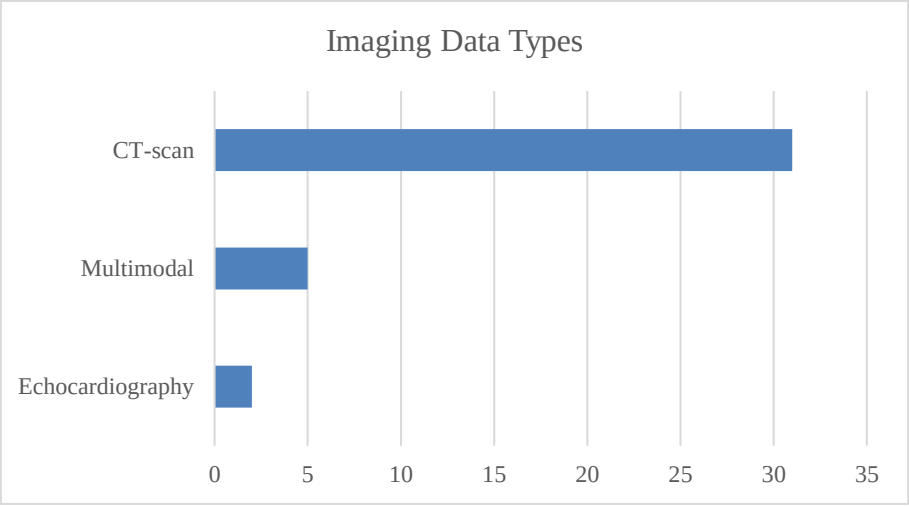


Figure 5 - Imaging data types used on analyzed papers

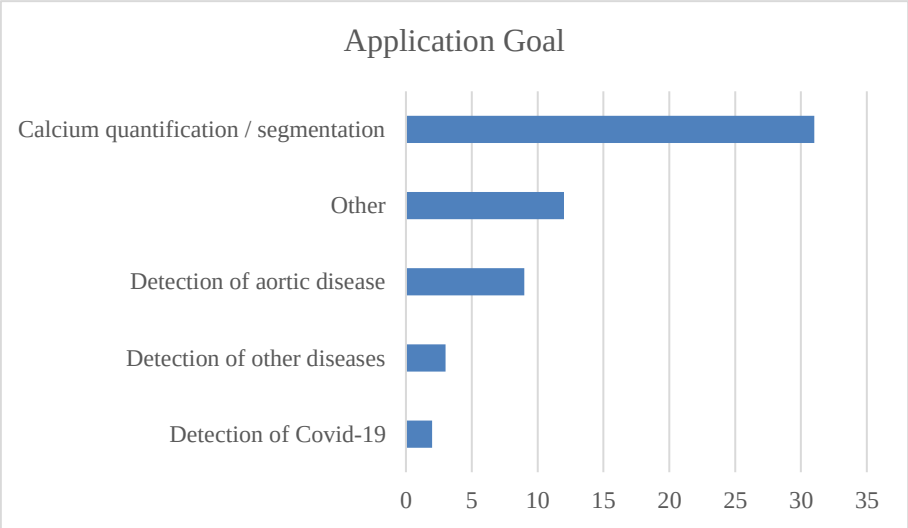


Figure 6 - Application goals of the analyzed papers

2.3. Related Work

The primary goal of this systematic review is to discover previous developments using AI for calcium detection on cardiac imaging, so the references were arranged according to the topics they cover, as shown on Table 2.

Deep Learning on Cardiac Images

DL is the most common popular subject, present in 95% of the documents included. There are some important advantages to using DL models in cardiac imaging. According to the state of the art, object detection, image segmentation and prediction are the most applied CV methods applied in health context [25].

In this review it is notable that CNNs have been applied to a variety of task such as (1) locating a heart in a CT image or (2) detecting calcium structures using image segmentation, where the objective is to produce a "filtered" image that only contains the desired object, as shown on Figure 7 from [25].

Table 2 - Topics analysis from the systematic review

Scope	Topic	References	# Docs	% Docs
ML Techniques	Convolutional NNs	[3], [5], [10], [14]–[19], [25]–[46]	31	82%
	Fully connected NNs	[14], [16], [19], [26]–[28], [39], [42]	8	21%
	Recurrent NNs	[28], [29], [39]	3	8%
	Other ML techniques	[10], [34], [44], [47]–[53]	10	26%
Imaging Data Types	CT-Scan	[3], [5], [14]–[19], [25]–[28], [30]–[33], [35]–[38], [40]–[43], [45], [47], [49]–[53]	31	82%
	Echocardiography	[10], [48]	2	5%
	Multimodal	[29], [34], [39], [44], [46]	5	13%
Application Goal	Calcium quantification / segmentation	[3], [5], [14], [16], [18], [19], [25]–[28], [30]–[37], [39]–[46], [49]–[53]	31	82%
	Detection of aortic diseases	[10], [25], [28], [34], [39], [44], [46]–[48]	9	24%
	Detection of Covid-19	[46], [49]	2	5%
	Detection of other diseases	[10], [34], [44]	3	8%
	Other goals	[10], [14], [15], [17], [25], [29], [33], [34], [38], [39], [44], [46]	12	32%

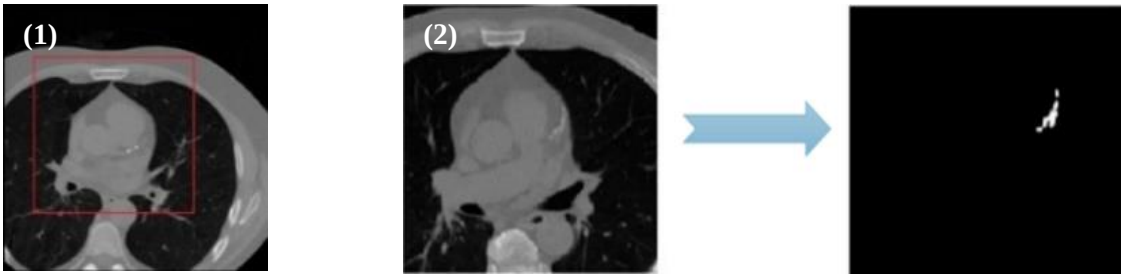


Figure 7 - Heart detection (1) and calcium segmentation (2) in CT image

DL is also useful for prediction, which might refer to regression or classification. Using cardiac images to identify a disease is a classification problem example. A binary classification model that divides images into "sick" or "not sick" classes (or calcified / not calcified in this specific case) and a multiclassification model where each disease represents a class and each image is given a diagnosis of disease "A," "B," or "C" are the two practical solutions for that context. Quantification of calcium score in aortic valve is an example of a regression problem, where the model predicts a continuous value.

Calcium Score

One of the greatest predictors of some complications like cardiovascular events, coronary heart disease or strokes is the presence of calcium in heart valves [17]. Aortic Stenosis and Aortic Sclerosis are two disorders that impact aortic valves. They are related to calcium formations in this heart valve, and they often have no symptoms. High coronary calcium scores and Covid-19 severity are also related, according to study [27], which used oxygenation measurements and CT image analysis.

The neural network used by the authors of [3], achieving 86% accuracy on cardiovascular risk classification, led to the conclusion that it might improve radiologists' workflow efficiency. Additionally, a model created on [13] demonstrates that by training a CNN architecture to repair blurring CT images, it is possible to increase calcium scoring accuracy, resulting in a significant reduction in assessment variances from 38% to 3.7%.

Workflow Efficiency Improvement

The implementation of a fully automated DL model, as mentioned in many studies, such as [3], [8], [15], [18], [21], [26], [37], [43], may lead to an increase in clinical workflow effectiveness. These automated models often show higher precisions than conventional techniques. According to the authors of study [20], ML-based processes have been found to be more accurate (80%) than manual processes done by physicians (60%) at predicting 5-year survival rates for patients with cardiovascular illnesses in the field of heart failure (HF).

Review [8] describes DL solutions for other classification problems, such as detecting some features like age and sex with a performance of 88% or distinguish two types of heart diseases from echocardiography images with a performance of 96.2%. Furthermore, deep neural networks are particularly useful for quantification problems. It would be beneficial for radiologists to avoid spending such a lot of time manually analyzing each screening to assess calcium score. This time-consuming and boring task can be done automatically by an algorithm, facilitating the process. The DL-based automated calcium quantification model referred to in [15] demonstrated high results when compared to manual assessment, confirming the practicality of its implementation.

On study [30] various tests were performed on calcium measurement. The training dataset returned a performance that was higher than manual scoring. Other experiments like quantifying the volume of epicardial adipose tissue (EAT) from CT scans on [32] has shown outstanding results, higher than 90%. Even though it cannot completely replace professionals, AI can significantly increase diagnoses' accuracy.

Radiation Exposure Reduction

The dangers associated to radiation exposure are a negative point of CT scanning, even though it is used to predict cardiovascular illnesses and death. This risk makes its usage constrained. According to the study [17], ultra-low dose (ULD)-CT techniques have been developed over time trying to reduce the effective radiation dose (ERD). Consequently, they increase noise, which lowers images' quality. Deep neural networks might provide a solution to this issue. Studies [17] and [44] have shown that DL algorithms can analyze "noisy" images produced by ULD-CT techniques, reducing noise, and enhancing the clarity of the images. This makes the use of low-radiation techniques more practical while protecting patients' health. An ULD-CT picture (A) can be changed into a noise-reduced CT image (B), as shown on Figure 8 from [44].

The use of echocardiogram imaging might represent another way to avoid radiation exposure. The use of sound waves instead of intrusive radiation works well on detecting a variety of heart and pulmonary diseases or assessing metrics like calcium score. Additionally, sound waves have great advantages such as their portability and minimal associated costs. Nevertheless, due to the lower image resolution, AI solutions in the field of echocardiography are currently scarce, which is an issue that can be solved with a higher volume of data [8].

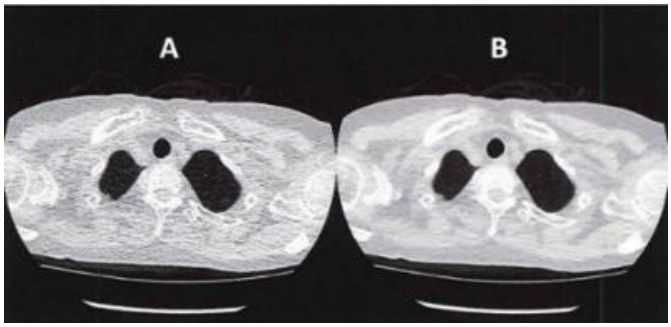


Figure 8 - Noise reduction in a CT-scan. The original image (A) is transformed in a noise-reducted image (B)

2.4. Summary

This literature review using PRISMA yielded 82 unique references, of which 38 met the goals of the eligible criteria, showing that there is still a lot to learn about this topic. DL, calcium score, and CT-scan are the key subjects discussed, primarily for calcium quantification objective. Particularly useful on cardiac imaging, including the identification of heart disorders and the detection and quantification of calcium, neural networks are often aborded. Only 13% of the Neural Networks were applied to echocardiography images; 87% used CT scans.

There are approaches presenting outstanding results, one of them outperforming manual processes [20], which shows that AI can significantly improve healthcare workflow efficiency. Using AI, boring and time-consuming tasks can be avoided, and diagnoses become quicker and more accurate.

Despite the radiation exposure risks to patients, CT scans are currently the medical imaging technique used on automation with DL, mostly because of their higher image resolution. Although avoiding this risk, echocardiography produces lower quality images, so this technique is less chosen. But since DL is available, this is no longer a problem, because it is possible to reduce noise and improve image quality.

Some developments in this area may close the current gap in fully automated methods for echocardiography images, increasing patients' monitoring frequency and reducing physicians' workload.

CHAPTER 3

Data Preparation

3.1. Data Understanding

The data selection process for this Dissertation was performed under the rules of a pre-signed confidentiality agreement to ensure compliance with data protection regulations while striving to create an appropriate dataset for the application of DL models. Through collaboration with Santa Maria's Hospital and Affidea clinics, a confidentiality agreement was executed to safeguard patients' data.

With the invaluable cooperation of a cardiology specialist from the Affidea clinic in Saldanha, Lisbon, who actively participated in all stages, echocardiography exams were exclusively selected from patients that consented to share their clinical data. As part of this dissertation´s work, these exams were manually extracted from an echocardiography machine at the Affidea clinic and saved in Dicom format [54]– a format specifically designed for medical files. Each of these files consists of two components: a set of images captured during the exam using sonography, and the associated metadata containing the patient's information. During the saving process, files were assigned anonymous names, such as "STUDY_000", "STUDY_001", and so forth, to ensure compliance with the general data protection regulations (GDPR).

Initially, 57 exams were copied, pre-processed, and utilized in modeling experiments. This initial dataset will be addressed as "Exam Set 1". Subsequently, 13 additional exams were incorporated in the initial dataset. Thus, a total of 70 echocardiography exams were filtered and subjected to the same processes outlined in the subsequent subsection. The final dataset will be addressed as "Exam Set 2".

The dataset distribution is unknown, since personal data like patients' gender, age or ethnicity were not provided because of the confidentiality agreement.

3.2. Pre-processing

From the initial selection of echocardiograms, a small subset was determined to be unrelated to heart conditions and subsequently excluded. Each examination comprises a set of images, typically totaling around ten.

The initial pre-processing procedure consisted of converting Dicom files into .jpeg image files, a more practical format that is both easier to handle and utilize. These images were stored with the examination ID incorporated into their respective names to facilitate identification.

Subsequently, a crucial step to ensure data protection and security was taken. Following the anonymization process, it was imperative to eliminate any personal data from each image. Echocardiography images inherently contain a header with the patient's name and other personal information. To address this concern, a Python script was developed to systematically scan through the images for removing the standardized-sized header. This procedure was conducted within the clinic immediately after the images were extracted from the echocardiography machine.

Throughout an echocardiography examination, the physician captures images of the heart from various angles, referred to as views. Within the acquired dataset, seven distinct types of views, detailed in Figure 9, were identified.

(A) Apical 2-chamber view
(B) Apical 3-chamber view
(C) Apical 4-chamber view
(D) Apical 5-chamber view
(E) Parasternal short-axis view
(F) Parasternal short-axis view of the left ventricle
(G) Parasternal long-axis view

Each view provides a focused perspective on a specific portion of the heart. Subsequently, every image was manually categorized based on its respective view, in collaboration with a clinical expert. For the designated purpose, it was determined to exclusively employ the parasternal short-axis views (Figure 9 - E), as these images feature the aortic valve, which constitutes the target of identification in the modeling phase. Following the selection of only this specific view, a dataset comprising 44 images from Exam Set 1 was established – being addressed as "Dataset 1". Subsequently, 17 images from Exam Set 2 – being addressed "Dataset 2" – were incorporated, resulting in a dataset of 61 images, all featuring the selected view.

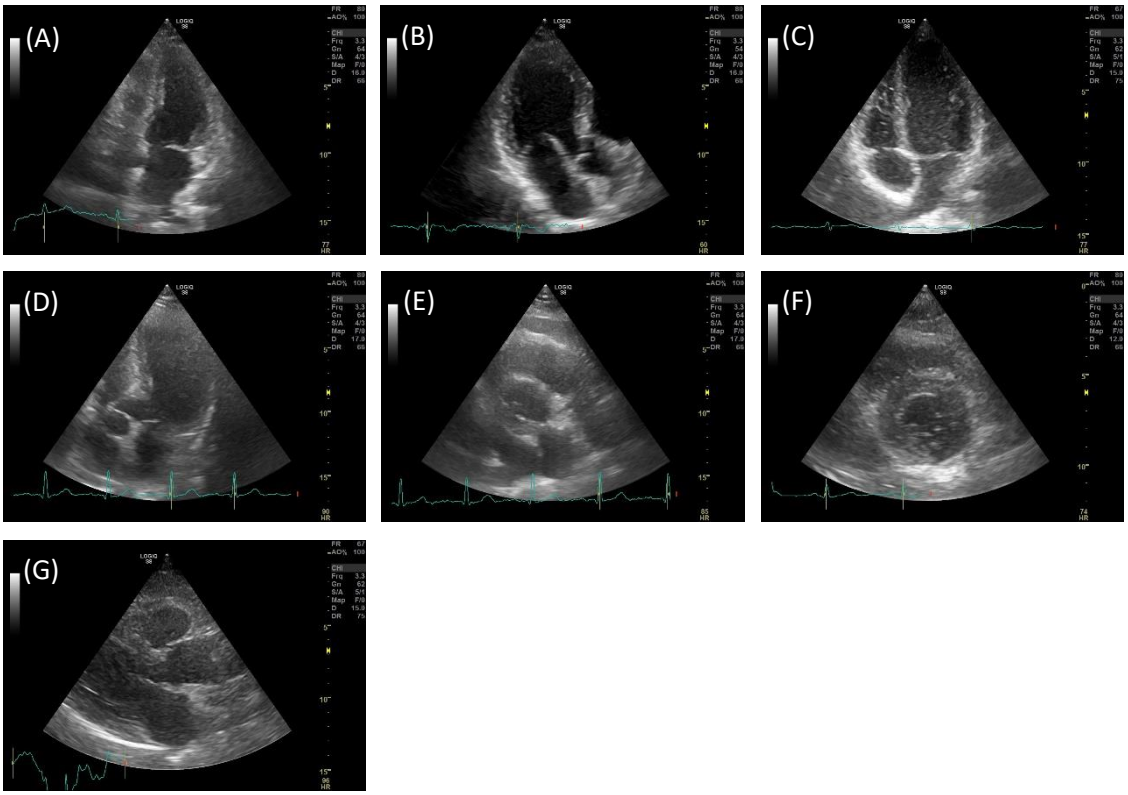


Figure 9 - Different types of heart view on Echocardiography: (A) Apical 2-chamber view; (B) Apical 3-chamber view; (C) Apical 4-chamber view; (D) Apical 5-chamber view; (E) Parasternal short-axis view; (F) Parasternal short-axis view of the left ventricle; (G) Parasternal long-axis view

The subsequent stage, illustrated in Figure 10, consisted of extracting and normalizing the dimensions of interest regions, using Python image processing tools. This process aimed to simplify the images by eliminating useless details that are automatically inserted on the images by the capturing device (which would otherwise be fed into the Neural Networks as noise).

(A) **Original image.** Echocardiography captures have a specific shape which looks like a cone. The region of interest is inside that cone, so the objective is to identify the cone and crop image according to that, filling all the space around with black.

(B) **Contours identification.** A Python script was built to identify the contours of the cone, with the use of functions from OpenCv library [55]that detect straight lines and circles.

(C) **Interest region selection.** A binary mask was created to select the area inside the three lines, which corresponds to the useful part of the image. Everything outside of those boundaries was set to black.

(D) **Edges crop.** Since the black area around the cone in the obtained images is also useless, the interest region was cropped to minimize the unnecessary black area.

(E) **Image crop to specific format**. All the images were manually cropped to 640×640 pixels dimensions (maintaining the aortic valve zone, which is the interest region). This standardization of images dimensions was taken to avoid images deformation, and consequently features loss. DL models require a specific input shape for images, and therefore a simple automatic resize to this input shape could potentially change the aspect ratio of the images, affecting the results. The dimension 640×640 pixels was chosen because it matches the input shape of the object detection model used in the experiments (YOLOv5 [56]).

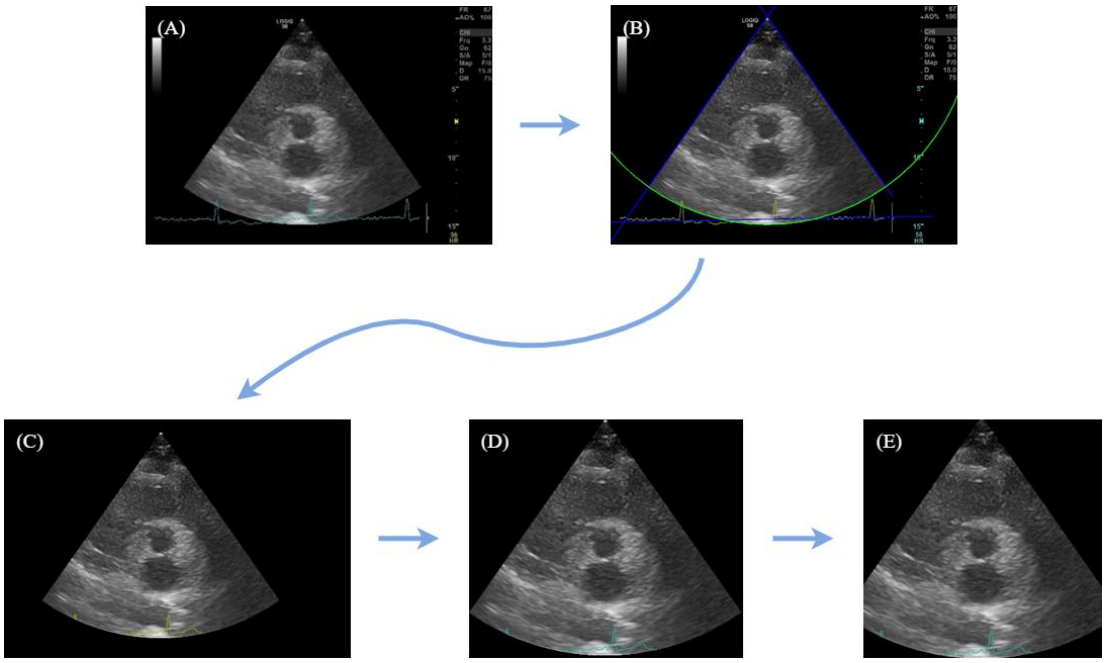


Figure 10 - Image transformation process

3.3. Annotations

In order to answer the first goal, it is necessary to train an object detector. To accomplish this, it is essential to have annotated images, i.e., the objects must be identified by a bounding box that indicates the object location in the image. With annotated images, a model is able to extract important features in its training phase and its performance can be evaluated in the testing phase by comparing its predictions with the annotations for the real object location.

This step was performed using MAXQDA [57], a software for data analysis with which is possible to manually draw bounding boxes and save all the annotations as the rectangle coordinates for each image in a .csv file. This process was applied to all 61 images selected for modeling phase. Figure 11 illustrates this process for one example image.

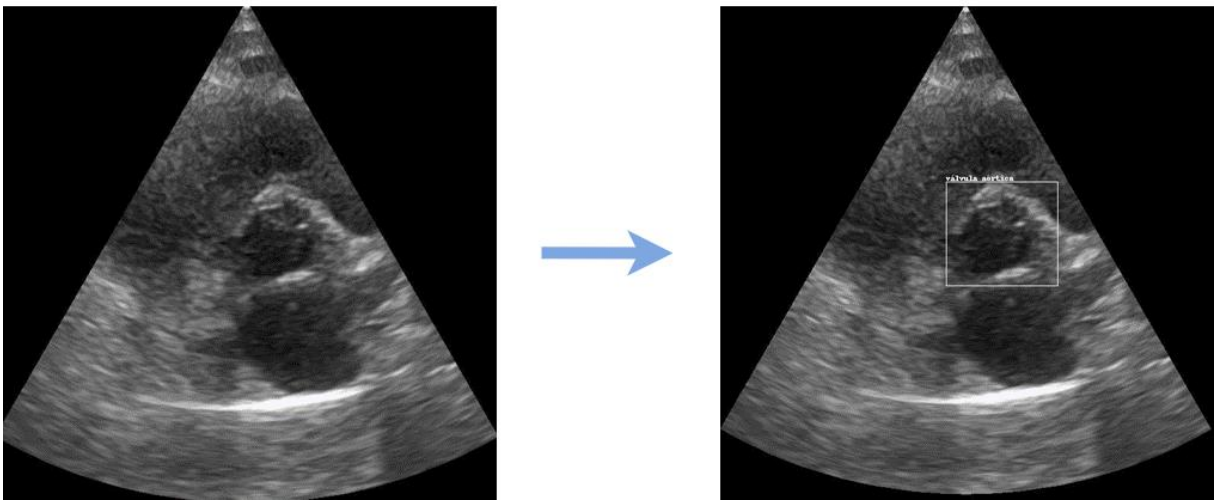


Figure 11 - Image annotation process

3.4. Data Augmentation

As the minimum dataset size to perform an object detector is usually around 150-500 images [58], the dataset obtained from Affidea with only 61 images is too small to achieve satisfactory results in the modeling phase. A commonly employed strategy to compensate for limited data availability is the application of image augmentation. This involves employing image transformation techniques to generate new images based on variations of the original ones, thereby expanding the dataset and introducing diversity for the model to learn from [59].

In this case, 4 types of data augmentation techniques were applied, to create a larger dataset before initializing the model train. The methods used were translation, rotation, zoom and gamma contrast, using a Python library called imgaug [60], which also transforms annotations directly, generating the corresponding object location in the augmented image version. In Figure 12, the images obtained from an original one by using each one of these techniques are illustrated.

(A) Translation. Translation percentage set to 0.2 and -0.2.

(B) Zoom. Image scale set to 0.5 and 2.

(C) Rotation. Rotation angle set to 10º and -10º.

(D) Gamma contrast. Contrast set to 0.5 and 1.2.

The rotation threshold was chosen according to the medical specialist's advice, justified by the exam procedure, where ultrasound position may vary about generally 10 degrees clockwise or counterclockwise from technician to technician. The other thresholds were chosen considering that the valve must be entirely inside the image after the transformation.

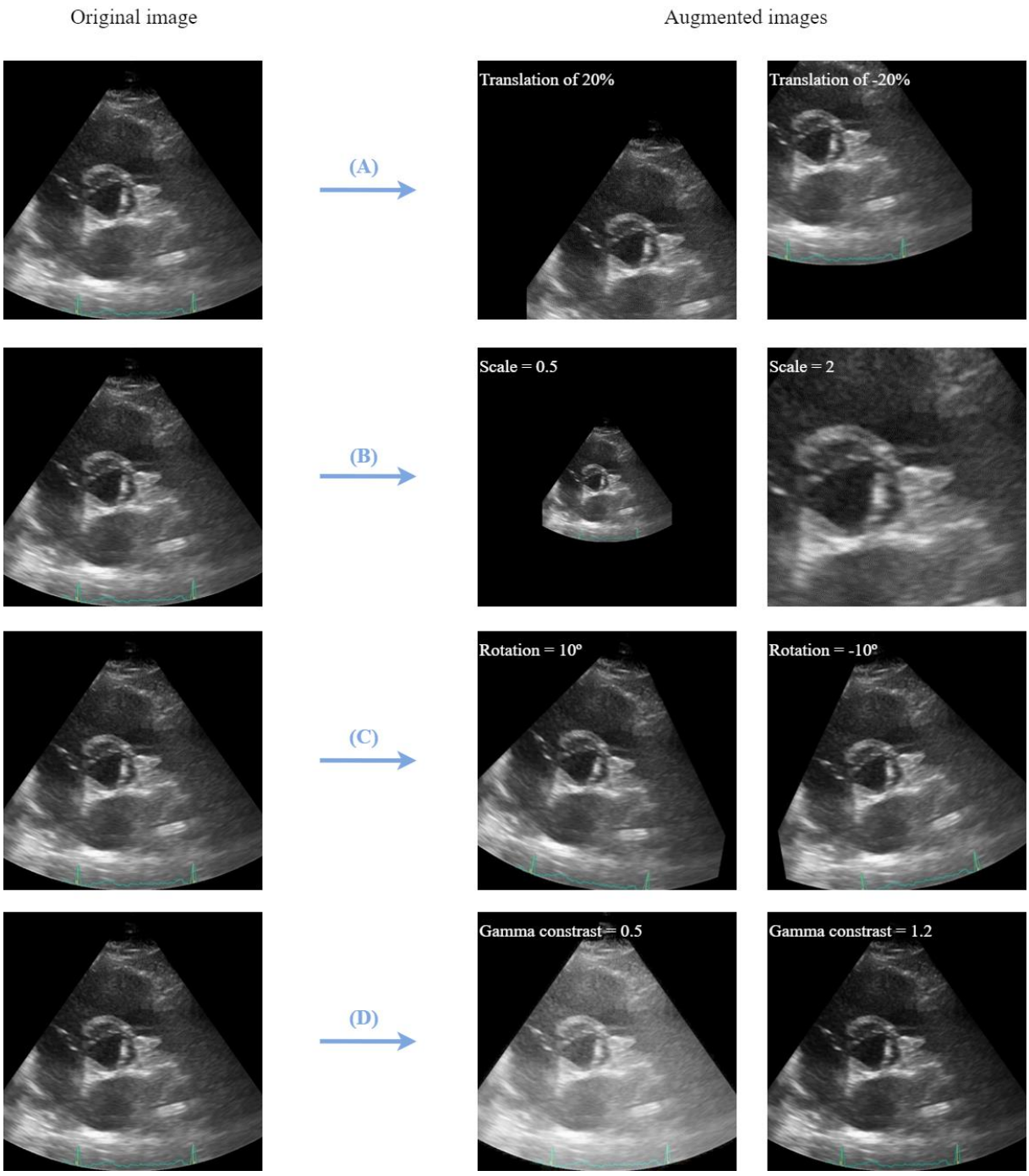


Figure 12 - Data augmentation process. (A) Translation; (B) Zoom; (C) Rotation; (D) Gama contrast

CHAPTER 4

Modeling

In the modeling phase of this dissertation, the main purpose is to build a framework for detection of AS in echocardiography images, by detecting calcium structures in aortic valve. To achieve the proposed goal, the development of two separate algorithms was necessary, as illustrated in Figure 13: (1) an object detector whose goal is to detect the aortic valve in each image, after training it with the created dataset; (2) a ML-based algorithm that classifies the image of the aortic valve according to the presence of calcium. For the second algorithm, a comparison between a DL approach and a heuristic classification was done, showing advantage on the use of DL models.

The simplest approach for this problem would use only the classification part, training the classifier with the entire original images. It would simplify the developments, but the expected results were very low, considering the small dataset size. Adding an object detector before the classification step, the classifier will be trained using only the area of interest, which is aortic valve. This makes it easier for the model to extract key features and to perform better.

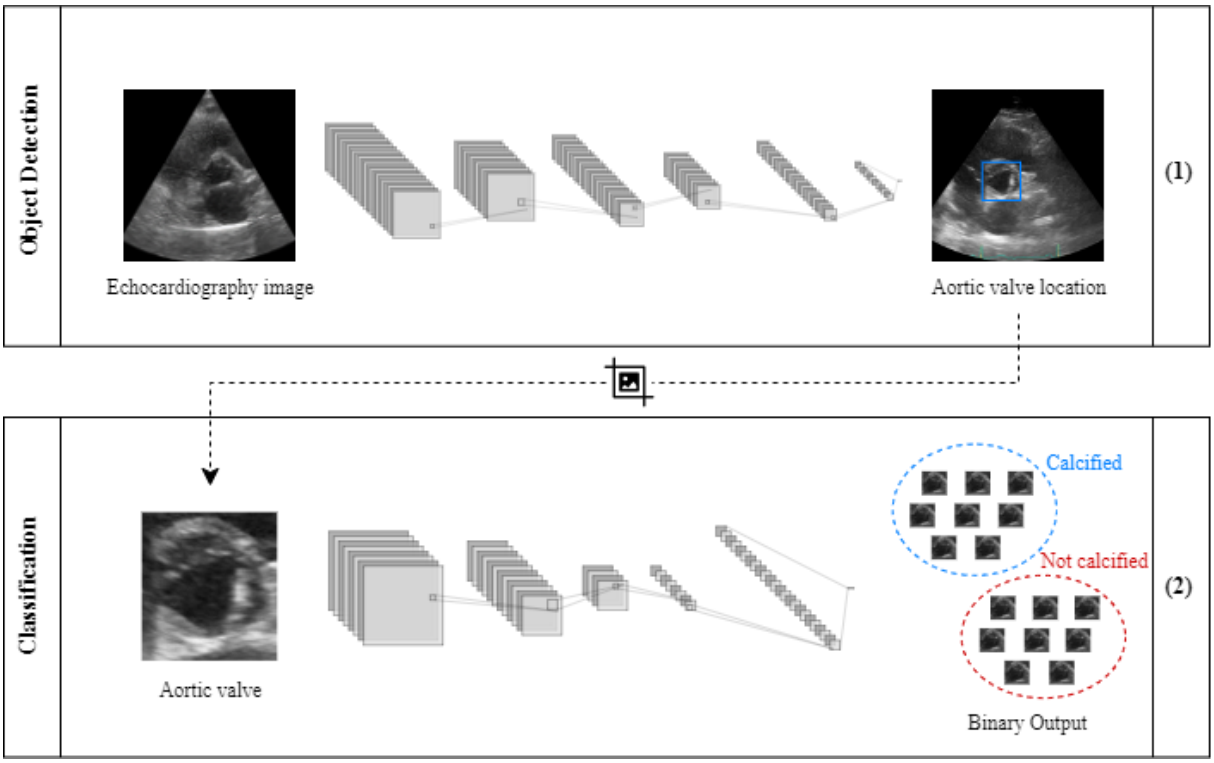


Figure 13 - Framework for aortic valve detection and Aortic Stenosis classification

4.1. Aortic Valve Detection

For the first algorithm, being the goal to automatize the identification of an object from an image (in this case the object is aortic valve), it was necessary to train an object detector. This CV technique consists of identifying objects within an image, by means of providing the bounding boxes around the detected objects.

4.1.1. Object Detector Modeling

There are two main ways to train an object detector, using DL [61]: (1) using Transfer Learning to apply a pre-trained model; (2) training a Neural Network from scratch, using a custom dataset, and manually defining weights and layers to start.

(1) Transfer Learning can be a faster method and achieve better results, as it uses a model that has already been trained with a huge quantity of data.

(2) Training from scratch requires a large amount of labeled data and typically requires more time to be trained.

The approach (1) was followed in this work, motivated by the lack of available images. The obtained dataset with 61 images would not be enough to build and train a model from scratch. As represented in Figure 14, a pre-trained model, previously trained with a large set of data called, was selected and its weights were used to initialize the NN. Then the next layers were trained using custom data – the echocardiography dataset with 61 images.

There are many useful libraries to work with CV, specifically with object detection, such as TensorFlow [62], Keras [63], PyTorch [64] and OpenCV [55]. To select a pre-trained model, it is important to consider the size and complexity of the selected dataset. From the variety of available models, YOLO's version 5 (YOLOv5) was chosen for demonstrating consistent performance, representing a turning point for YOLO's model versions, as it is the first one built on the PyTorch framework [56]. Since working with a small set of images, the YOLOv5s variation – which is designed for small datasets – was selected in this case.

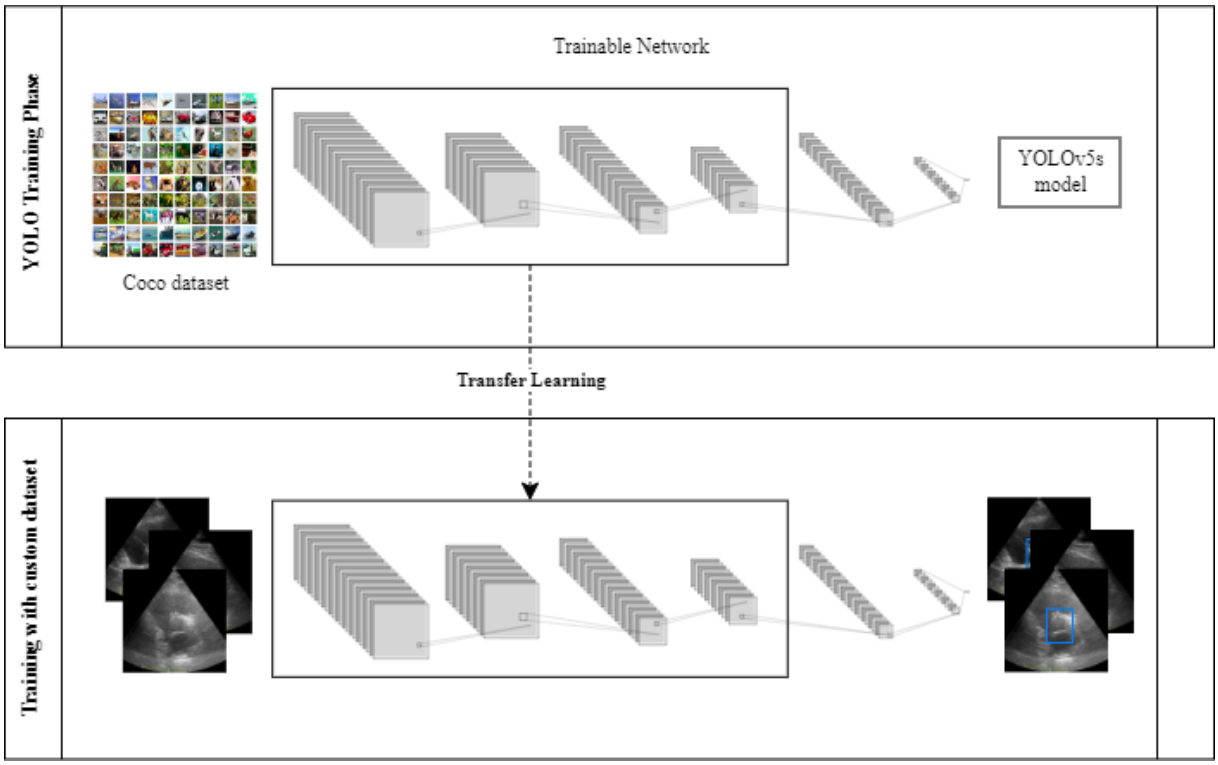


Figure 14 - Object detector with Transfer Learning for aortic valve detection

4.1.2. Performance Metrics

To evaluate the object detector performance, three metrics were calculated for each model. These metrics, based on the comparison between the objects detected by the model and the ground truth, are commonly used in object detection problems and are based on the outcome of the model's predictions [65]. To calculate those metrics, it is important to define the following concepts:

Intersection over Union (IoU) is an object detection measure used to evaluate if a detection is correct or not, by analyzing the overlap between the bounding boxes detected by the model and the bounding boxes in the ground-truth, calculated by

$$IoU = \frac{Area\ of\ intersection}{Area\ of\ union}, \qquad (1)$$

as illustrated in Figure 15.

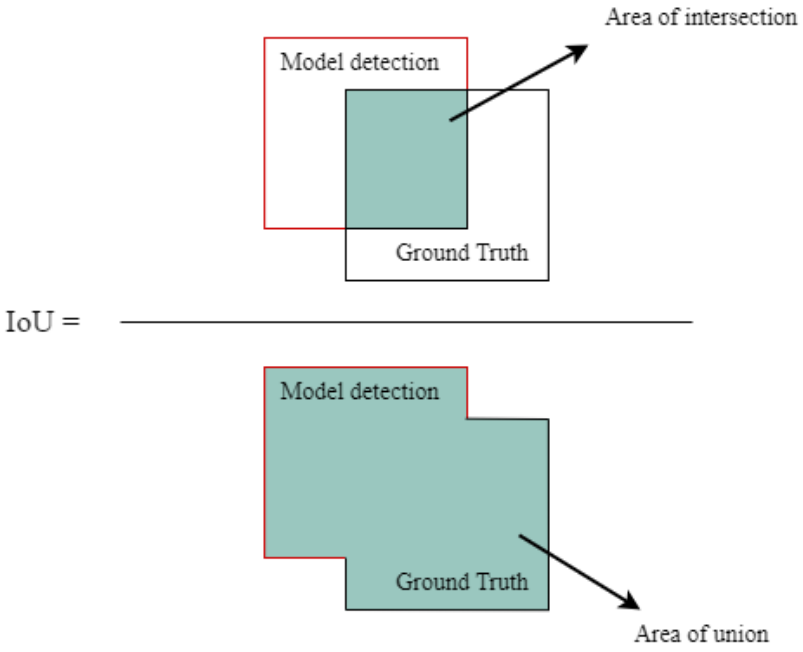


Figure 15 - IoU calculation

It is represented by a value between 0 and 1, where 0 represents no overlap and 1 represents perfect detection. To distinguish between valid or invalid detections, a threshold value, α, was defined to 0.2. After trying some thresholds, this value showed better results [66].

- TP (True Positives) → Represents the number of objects detected by the model, meeting a IoU value ≥ α.
- FP (False Positives) → Represents the number of objects detected by the model, meeting a IoU value < α.
- FN (False Negatives) → Represents the number of actual objects that were not identified by the model.
- TN (True Negatives) → Not calculated for this model as there was only one object class.

Based on these definitions, three metrics were calculated and used to evaluate models' performance:

- Precision → Number of correct predictions over the positive predictions made by the model, computed as

$$Precision = \frac{TP}{TP+FP}; \quad (2)$$

- Recall → Number of correct predictions over the actual positive cases, calculated as

$$Recall = \frac{TP}{TP+FN}; \quad (3)$$

- F1-score → A weighted average between precision and recall, computed as

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (4)$$

4.1.3. Experimental Procedures

Dataset 1 – 43 images

The first model was trained with the first set of images copied from the echocardiography machine, containing 43 images. The following parameters were defined when creating the NN:

- Input shape → (640, 640, 3);
- Number of batches → 32;
- Number of epochs → 75;
- Dataset split → 80% for train + 10% for validation + 10% for test.

For hyperparameters specification, a file created by default in YOLO folder was used. Among other parameters, this file defines if data augmentation is used or not. In this case, data augmentation was set to be done, applying a translation and a zoom on each image during the train. To understand how much impact the data selection can have, a K-fold Cross Validation [67] was performed, setting K to 10. Data was split into 10 equal parts, generating 10 different combinations train/test. The model was trained 10 times, using each subset as a test set once. The results are presented in Table 3.

Table 3 - Aortic valve detection results using Dataset 1

Experiment	Dataset size	Precision	Recall	F1-Score	Test Set
1	43	0.8	1	0.89	Test Set 1
2	43	1	0.75	0.86	Test Set 2
3	43	1	0.75	0.86	Test Set 3
4	43	0.67	0.5	0.57	Test Set 4
5	43	0.5	1	0.67	Test Set 5
6	43	0.67	1	0.8	Test Set 6
7	43	0.57	1	0.73	Test Set 7
8	43	0.43	0.75	0.55	Test Set 8
9	43	-	0	-	Test Set 9
10	43	1	0.75	0.86	Test Set 10
Global	**43**	**0.74**	**0.75**	**0.75**	**K-fold Cross Validation**

From the observations, experiment 9 stands out as having the worst performance, with 0 detections, leading to a recall value of 0 and precision and f1-score that cannot be calculated. Analyzing the test set for this model, shown in Figure 16, the images are too dark and/or contain too much noise, increasing the difficulty for the valve identification.

The other experiments registered some variations, but not very significantly. It is highlightable the first experiment, achieving 80% of precision and 100% of recall, meaning that all the existent valves were correctly detected, and a small part of model detections were false.

Globally, this model achieved around 75% for the three models, which is a positive and promising result.

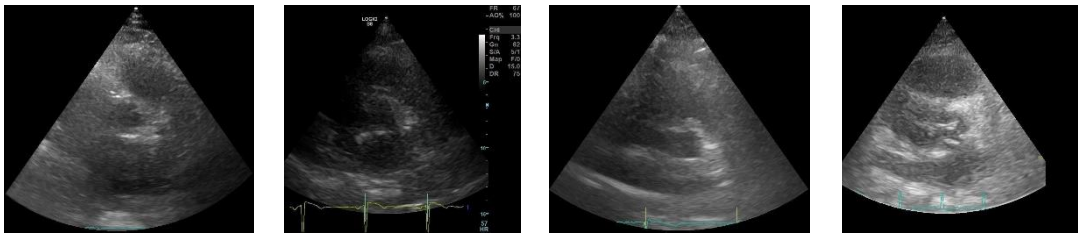


Figure 16 - Test set of experiment 9, where the model could not detect aortic valve

Dataset 2 – 61 images

In a second approach, all the images had been extracted from the clinic and pre-processed, so the full dataset, containing 61 original images, was used. The same specifications as in the first approach were applied:

- Input shape → (640, 640, 3)
- Number of batches → 32
- Number of epochs → 75
- Dataset split → 80% for train + 10% for validation + 10% for test

Then the images resulting from 4 types of data augmentation techniques, described in Section 3.4., were gradually included, to train the object detector with a larger dataset. Each data augmentation technique was included individually (Experiments 2 to 5), then a dataset using all the techniques simultaneously was used (Experiment 6), and finally a model using the original images and the data augmentation images resulting from the two techniques that had the best individually performances – translations and rotations – was trained (Experiment 7). All the results are summarized in Table 4.

When using dataset 2, with 61 original images, the observable performance results are different from the registered for dataset 1. The first experiment, without data augmentation, achieved a low performance. Although having a 100% precision, it registered only 26% of recall, which means that all the model predictions were correct, but it could not detect all the actual objects (only 26%).

Table 4 – Object detection results using Dataset 2

Experiment	Dataset size	Precision	Recall	F1-Score	Description
1	61	1.00	0.29	0.44	Original dataset
2	183	1.00	0.81	0.89	Data Augmentation - zoom
3	183	1.00	0.95	0.98	Data Augmentation - translation
4	183	0.95	1.00	0.98	Data Augmentation - rotation
5	183	0.95	0.90	0.93	Data Augmentation - gamma contrast
6	549	0.91	0.97	0.94	Data Augmentation - zoom, translation, rotation and gamma contrast
7	305	0.93	1.00	0.97	Data Augmentation - translation and rotation

By including the image sets obtained from each data augmentation technique, the results increased significantly, with highlight for translation and rotation (experiments 3 and 4). The use of translation for enlarging the dataset increased the precision to 100% and recall to 95%, meaning that all the model predictions were correct, and almost all the actual locations of the valve were identified (only 5 % were not found). Rotation, on the other hand, achieved 95% of precision and 100% of recall, which means that all the actual valves were correctly located by the model, and 5% of its predictions were false. In health context, recall usually gets more importance, as it is better to get a false positive than a false negative, i.e., is it better to predict a disease presence in an actual health patient than not identifying the disease [65]. In this case, although the objective is to identify the aortic valve, in the next phase the purpose is to classify this valve as calcified or not, to detect AS cases. So, between experiment 3 and 4, giving priority to recall, it would be preferable to use rotation technique (experiment 4) for deployment, as it achieved 100% recall.

When including all the images obtained with the four data augmentation techniques, the training time increased significantly, because of the dataset size that increased images processing time, and the results decreased to 91% of precision and 97% of recall. Experiment 4 maintains an advantage over this one. Lastly, a training was performed including the two techniques that showed higher results individually – translation and rotation – expecting to get even higher performance. Recall result was 100% again, but precision decreased to 93%, so it has no advantages, when compared to the use of only rotation.

The analysis of models training is also important to better understand the results obtained. When trained with the original dataset, with 61 images, the object detector yielded the confusion matrix represented in Figure 7. As it can be seen, only 2 of 7 aortic valves were correctly identified, justifying the 29% recall. Although having 100% precision, in this case this is not a good indicator because it means that the only 2 aortic valves identified by the model were correct.

In the loss and metrics curves represented in Figure 18, it is notable the inconsistency of the model during train. The loss decreases along the training, but always oscillating. Precision and recall increase with train, but have the same unstable behaviour.

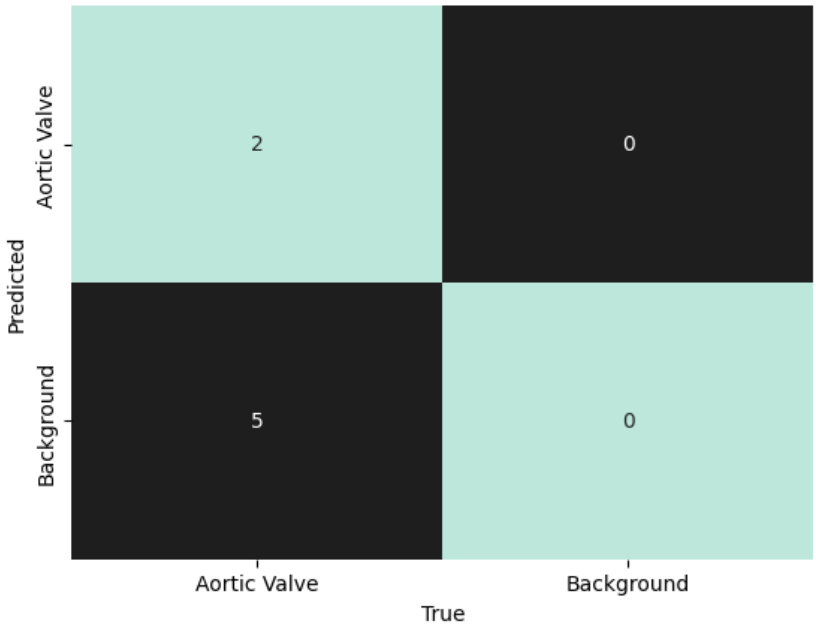


Figure 17 - Confusion matrix for object detection with original dataset

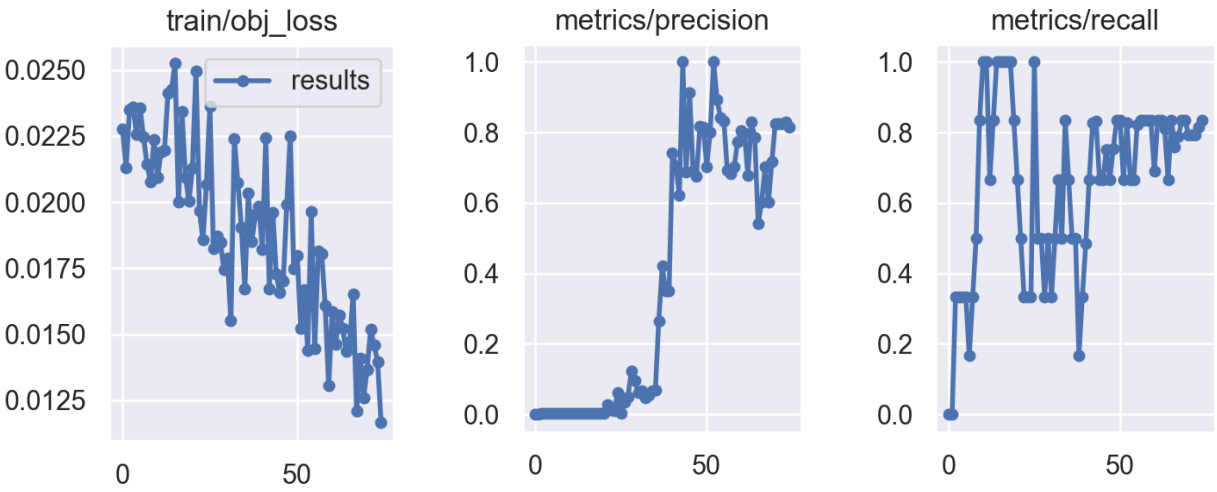


Figure 18 - Loss and metrics curves for object detection with original dataset

In Figure 219, the application of the model trained with the original data is demonstrated, based on some example images. The model was only capable of identify 1 aortic valve out of the 4 images tested, which indicates that the neural network did not extract enough information from the training, not being capable of adapting to new objects.

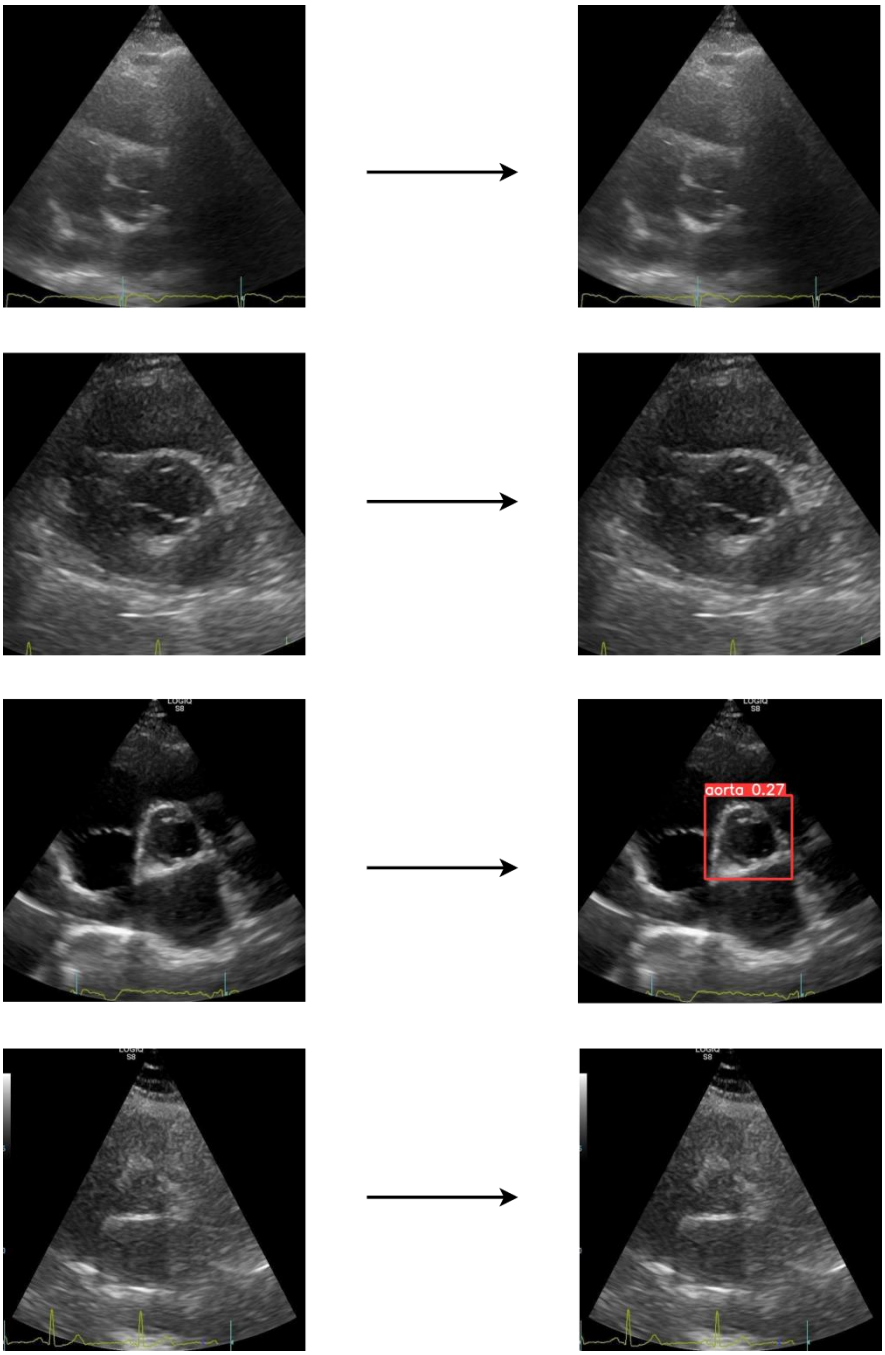


Figure 179 - Aortic valve detection with the original dataset - examples

Comparing with the best model – with rotated images – the behaviour during train is different. With this larger dataset, the confusion matrix obtained by the object detector, represented in Figure , shows that the 21 real aortic valves were correctly identified. The model only registered 1 False Positive, demonstrating an almost excellent performance.

Also the loss and metrics curves, represented in Figure 21, describe a more consistent training. The loss decreases in a constant way along the training, and precision increases in the same way. Recall is unstable in the beginning of the train, but in the end becomes consistent.

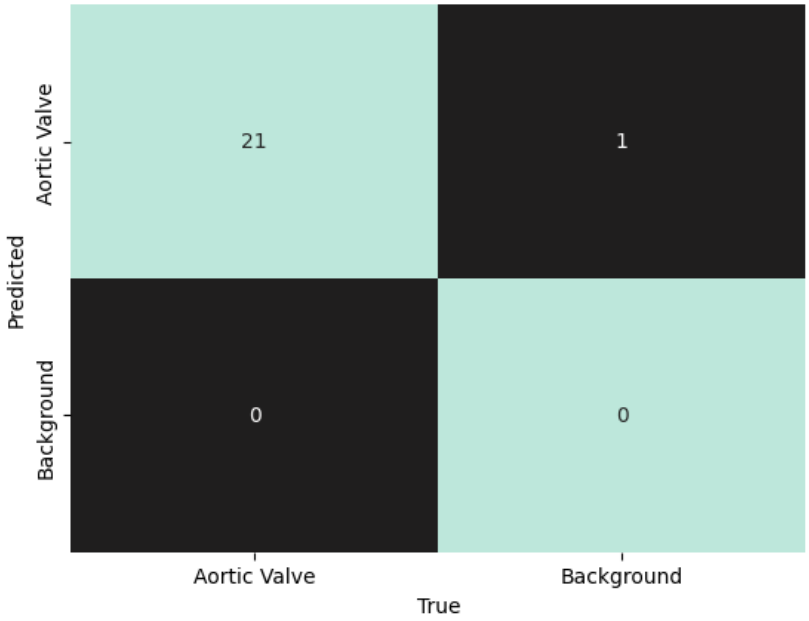


Figure 20 - Confusion matrix for object detection with data augmentation (rotations)

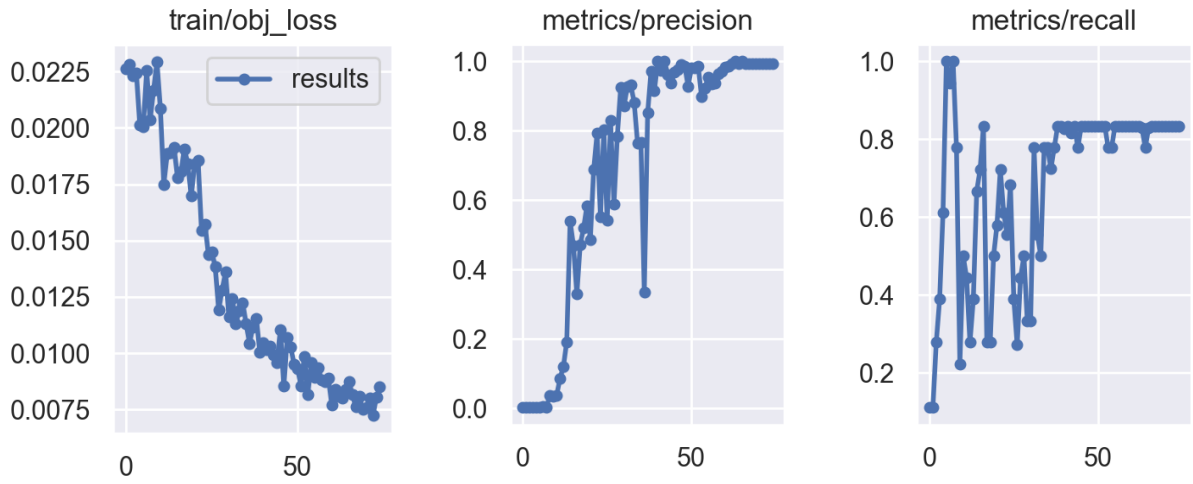


Figure 21 - Loss and metrics curves for object detection with data augmentation (rotations)

In Figure 22, the application of the model considered the best is demonstrated, based on some example images. As it can be seen, the model is capable to identify the aortic valve in various types of images, since rotated ones – like the first two – to images with some noise – like the last one.

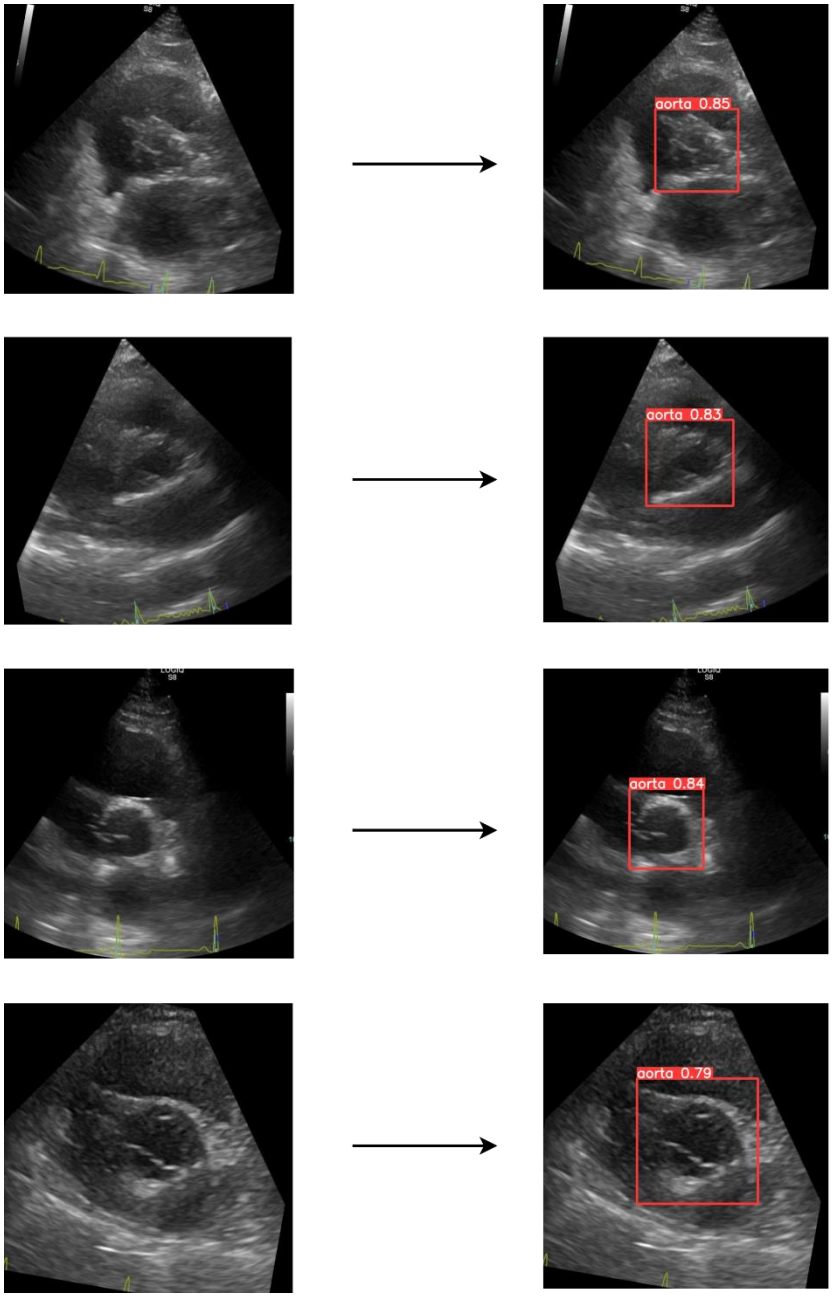


Figure 182 - Aortic valve detection with the best model - examples

4.2. Aortic Stenosis Classification

One approach to automatize diagnosis of AS is to use a DL-based classifier that distinguishes between calcified or not calcified valves, in echocardiography images. Like the object detector mentioned in Subsection 4.1.1., the classifier can both be applied from a pre-trained model using Transfer Learning or be built from scratch. There are many classification models available, such as EfficientNet [68], MobileNet [69] or ResNet [70], that will be applied and compared. A heuristic approach will be also tested, as a baseline method, to evaluate the advantages on using DL.

4.2.1. Classifier Modeling

Regarding the development of the DL model for automatic classification, Transfer Learning was used, because it is a great option when using a small dataset, like in the presented context. Various existent models can be used to initialize the network's weights, and the model must be chosen according to the data to apply it, to get the best performance. Three models were tested - EfficientNet, MobileNet and ResNet.

Figure 223 represents the use of Transfer Learning with the EfficientNet model, which is similar when applied to other classification models. As can be seen, these models are trained using ImageNet dataset, learning to classify images into categories. When applied to a custom dataset, the first layer will have the pre-trained weights and the next layers are trained with the new data.

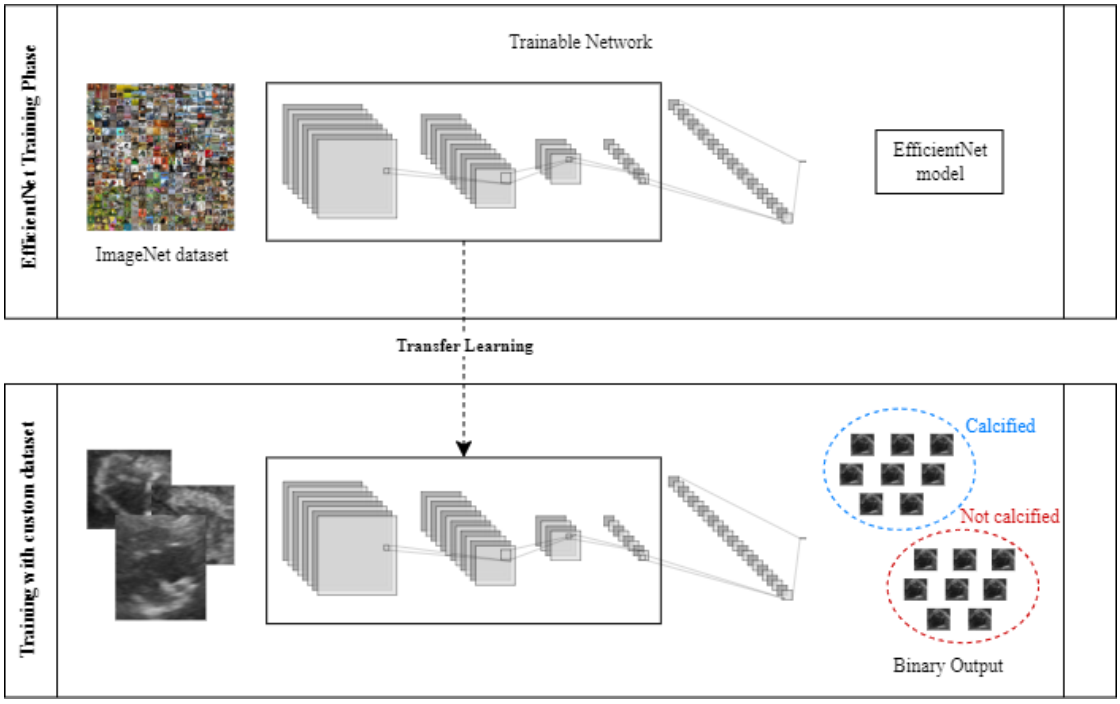


Figure 193 - Classification model with Transfer Learning for Aortic Stenosis classification

4.2.2. Performance Metrics

In classification context, the metrics used for performance evaluation are accuracy, precision, recall, f1-score, ROC AUC, and log-loss [71]. The following definitions were considered:

- TP (True Positives) → Represents the number of correct identifications of AS cases by the model.
- FP (False Positives) → Represents the number of health patients classified as having AS by the model.
- FN (False Negatives) → Represents the number of AS cases classified as healthy by the model.
- TN (True Negatives) → Represents the number of correct classifications of healthy patients.

Based on these concepts, the three performance metrics used in object detection, and an additional one – accuracy – were calculated and used to evaluate models' performance:

- Accuracy → Number of correct predictions over the total evaluations, calculated as

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}. \quad (5)$$

4.2.3. Experimental Procedures

Baseline Heuristic Approach

Based on [20], the first approach was performed using a heuristic classification. This technique consisted of binarizing images and searching for white pixels (RGB = (255, 255, 255)). If the image has at least one white pixel, the valve is considered calcified. This process is represented in Figure 224.



Figure 204 - Heuristic classification process

The binarization process was manual, to try different thresholds and find the best one. Four different thresholds were tested – 190, 200, 210 and 220 – which means that all the pixels with higher value than the threshold would be set to 255 (white) and the remaining ones would be set to 0 (black). The results obtained for these heuristic approaches are summarized in

Table 5.

Table 5 - Aortic Stenosis classification results for heuristic experiments

Experiment	Dataset size	Accuracy	Precision	Recall	F1-Score	Binarization threshold
1	183	0.42	0.33	0.58	0.42	190
2	183	0.52	0.35	0.39	0.37	200
3	183	0.58	0.40	0.32	0.36	210
4	183	0.39	0.33	0.64	0.43	220

From the four heuristic approaches, the models registered accuracy values be-tween 39% and 58%, precision between 33% and 40% and recall between 32% and 64%, noting that when one of these metrics increases, the other two decreases (for instance, comparing experiments 3 and 4). In this case, as it is preferable to have higher recall than accuracy or precision [72], to minimize the False Negative cases, i.e., AS cases that are not detected, the best model to choose would be 4, using a binarization thresh-old of 220. However, this model's accuracy and precision are quite low, which means that there are many False Positives, i.e., the model identifies too many AS cases where the patients are healthy.

Initial DL Experiments

Attempting to get better results, a DL-based approach was followed, consisting of building a Neural Network with the structure shown in Figure 225. The first layer's set – responsible for feature extraction and composed by convolutional and pooling layers – uses the pretrained models previously mentioned. The following layer is a dense layer with 120 neurons, and a final output layer performing a binary classification meaning "calcified" or "not calcified". The model was trained using the full dataset, containing 61 images.

- Input shape → (224, 224, 3);
- Batch size → 1;
- Number of epochs → 50;
- Dataset split → 80% for train + 20% for test, then 10% of train was used for validation.

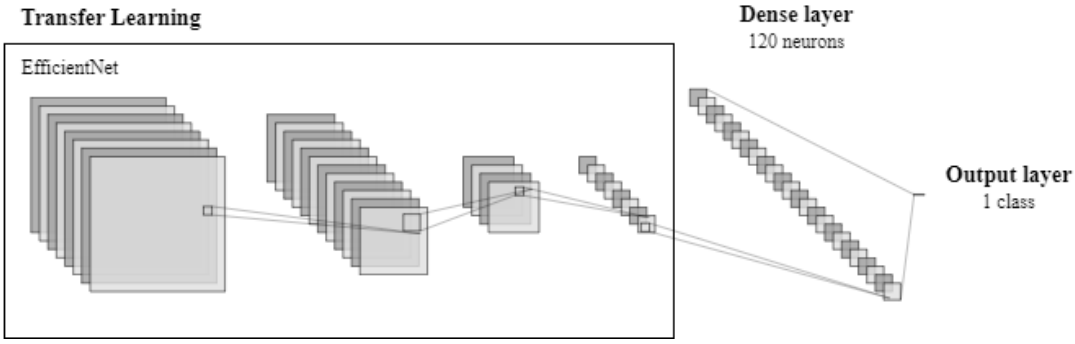


Figure 215 - Neural Network architecture

First, the images were entirely used to train the model, i.e., without cropping the valve zone. As result, the model predicted around 50% probability of being calcified, which is a sign that the model did not learn to differentiate between classes. There are studies referring that cropping images can improve the feature extraction process, because when training an image classification model, it tries to extract features from the entire image, which can become less efficient when having useless information [73]. In this case, images were cropped using the original bounding boxes, with a Python script that selected the area between the given coordinates for each image, obtaining the area of interest that contained aortic valve, as illustrated in Figure 26.

In the next step, for Transfer Learning application, three pre-trained models were used – EfficientNet, MobileNet (and MobileNetV3Small, which is a variation for smaller datasets) and ResNet50. All the three models returned the same output – all the images were classified in the majority class. This means that the network could not extract relevant features from the images, so it predicted the class that was more likely to be correct (the most common).

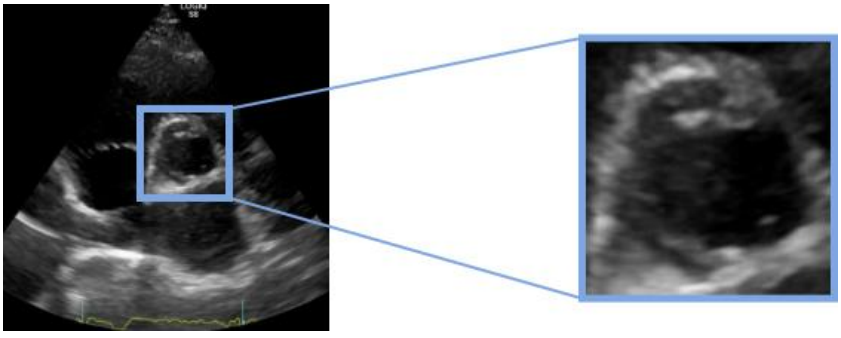


Figure 226 – Process of cropping the area of interest

Data Pre-Processing Experiments

Trying to improve the previous results, the dataset was adapted to avoid biased training of the classifiers. The dataset was modified using the following techniques:

- **Noise reduction.** Creation of an algorithm that scans all the pixels in an image and subtracts the minimum value to the entire image array. The objective is to turn the darker zones into pure black (RGB = (0, 0, 0)), as demonstrated in Figure 27.
- **Data balancing.** The original dataset was unbalanced between the two classes – 22 "calcified" images and 39 "not calcified" images, which can be a problem when the model learns to always predict the same class (the majority one) because the probability of it gets correct is higher [74]. To get around this problem, some images in the minority class were duplicated until the two classes become balanced.
- **Data augmentation – Rotations.** The rotated images created in data preparation phase, referred to in subsection 3.4., were included to enlarge the dataset. The other types of data augmentation were not adequate to this classification context. Zooms and translations would not work on images containing only the region of interest, and contrast variations would turn the dataset non-uniform.



Figure 237 - Process of noise reduction

The following experiments employed two pre-trained models – EfficientNetB0 [75] and MobileNetV3Small [76]. These models were chosen for their simplicity, which is preferrable when working with smaller datasets. More complex models tend to better explain the training data, possibly getting overfitted [77]. The original dataset was used, and two variations were tested – class balancing and data augmentation. The first one equalizes classes distribution and the second one augments the dataset by incorporating rotated images generated during the data preparation phase outlined in Subsection 2.4., were included. The obtained results are described in Table 6.

In the first experiment, where the original dataset was used for training, both pre-trained models achieved an accuracy of 54%, which is insufficient to draw conclusions. The second experiment, involving class balancing of the original dataset, yielded 0 TP's, indicating an inability to correctly classify images with calcified valves, leading to precision and recall values of 0. Thus, this NN is not promising as it fails on identification of unhealthy patients.

Using a larger dataset, performance improved. In experiment 3, rotated images were included without balancing the dataset. EfficientNetB0 returned inconclusive results, displaying 100% precision but only a 6% recall, suggesting a failure to identify the majority of truly calcified valves. In contrast, MobileNetV3Small could achieve more than 70% for all performance metrics, indicating better feature identification and generalization.

When balancing the augmented dataset, NN were trained with 234 images split into two balanced classes. Both pre-trained models had similar performances, achieving over 90% for all metrics. While EfficientNetB0 had consistently 96% for accuracy, precision and recall, MobileNetV3Small showed better precision than recall, remaining more unhealthy patients to be identified.

Table 6 - Aortic Stenosis classification results for experiments using different datasets

Experiment	Dataset	Dataset size	Pre-trained model	Accuracy	Precision	Recall	F1-Score
1	Original	61	EfficientNetB0	0.54	0.43	0.60	0.50
			MobileNetV3Small	0.54	0.33	0.20	0.25
2	Balanced	78	EfficientNetB0	0.31	0	0	-
			MobileNetV3Small	0.44	0	0	-
3	With data augmentation	183	EfficientNetB0	0.57	1.00	0.06	0.11
			MobileNetV3Small	0.84	0.92	0.71	0.80
4	Balanced with data augmentation	234	EfficientNetB0	0.96	0.96	0.96	0.96
			MobileNetV3Small	0.94	0.96	0.92	0.94

Experiment 4, yielding the most promising results, highlights the significance of class balancing and data augmentation to enhance model's performance. Subsequently, this dataset was used to train a NN, with variations in the pre-trained models used for Transfer Learning. Six pre-trained networks – EfficientNetB0, EfficientNetB4, Res-Net152V2, ResNet50V2, MobileNetV3Small and MobileNetV3Large – were trained, and the pooling parameter was adjusted to three values for each model – "None", "Max" and "Avg". Pooling parameter defines the feature extraction mode during Transfer Learning. When set to "None", the model's output mirrors that of the last convolutional layer, while "Max" and "Avg" apply maximum and average pooling to the last convolutional layer, respectively. All the experiments are summarized in Table 7.

For the majority of pre-trained models tested, setting the pooling parameter to "None" achieved better performance. This observation may be attributed to the model's challenges in feature extraction, particularly given the dataset's size. Training with a small dataset could limit the model's ability to extract information from images, and pooling application might further reduce available information.

Among the six pre-trained models used, EfficientNetB0 and MobileNetV3Large stand out with better performances. The first one achieved 96% for all the metrics, while the second one achieved a F1-score of 96%, with a 100% recall, indicating that all the unhealthy patients were identified. In general, simpler NNs demonstrated better performance, likely due to their training with a smaller dataset, requiring less complexity. In this context, considering the importance of recall, MobileNetV3Large with-out the use of map pooling may be considered the best approach.

Table 7 - Aortic Stenosis classification results for balanced and augmented dataset using different pre-trained models

Experiment	Pre-trained model	Map pooling	Accuracy	Precision	Recall	F1-Score
1	**EfficientNetB0**	**None**	**0.96**	**0.96**	**0.96**	**0.96**
2		Max	0.75	0.67	0.92	0.79
3		Avg	0.88	0.88	0.88	0.88
4	EfficientNetB4	None	0.92	1.00	0.83	0.91
5		Max	0.86	0.90	0.79	0.84
6		Avg	0.79	0.89	0.67	0.76
7	ResNet152V2	None	0.85	0.84	0.88	0.86
8		Max	0.77	0.76	0.79	0.78
9		Avg	0.85	1.00	0.71	0.83
10	ResNet50V2	None	0.71	0.63	1.00	0.77
11		Max	0.81	0.8	0.83	0.82
12		Avg	0.77	0.88	0.63	0.73
13	MobileNetV3Small	None	0.94	0.96	0.92	0.94
14		Max	0.69	0.74	0.58	0.65
15		Avg	0.85	0.9	0.79	0.84
16	**MobileNetV3Large**	**None**	**0.96**	**0.92**	**1.00**	**0.96**
17		Max	0.77	0.81	0.71	0.76
18		Avg	0.79	0.94	0.63	0.75

For a better interpretation of the results during train, confusion matrices, loss and accuracy curves were analysed. Highlighting one of the best models – EfficientNetB0 – the confusion matrix represented in Figure 28 shows that the model classifications are well distributed, and 46 out of 48 cases are in the correct class.

From the curves represented in Figure 29, it is observable that accuracy increases during train, in a fast and stable way. Training and validation curves are near each other, showing that the model is being adapted to the training data. Loss decreased in the beginning of the training, staying low until the end, which confirms that the model is being correctly adapted to the data, reducing lost information.

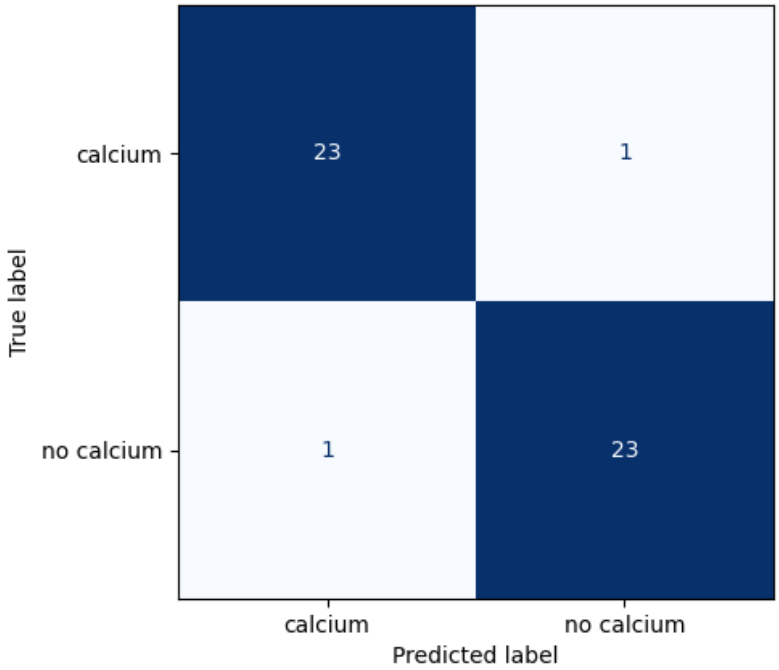


Figure 258 - Confusion matrix for image classification with EfficientNetB0

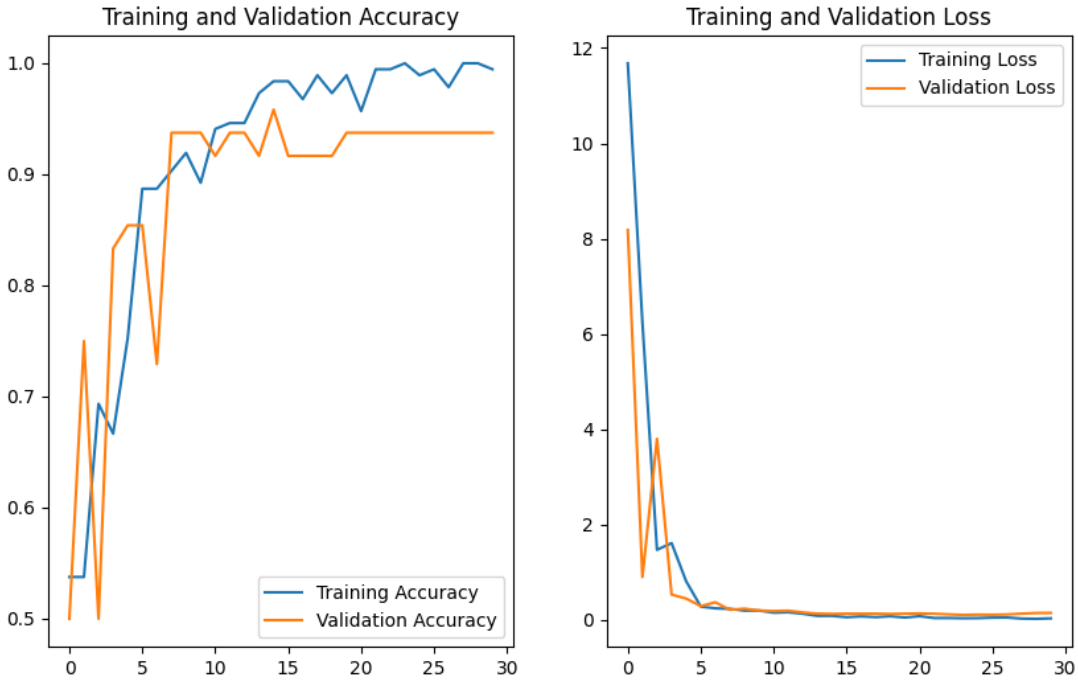


Figure 249 - Accuracy and loss curves for image classification with EfficientNetB0

4.3. Discussion

To identify Aortic Stenosis (AS) in echocardiography images, the study focused on two main goals: (1) detecting the aortic valve and (2) identifying the presence of calcium in the aortic valve.

In the initial phase, an object detector was constructed to locate the aortic valve in echocardiograms. Transfer Learning was employed from a pre-trained model due to the limited amount of available data. YOLOv5s was chosen as the pre-trained model for its reliability, particularly in handling small datasets.

However, the initial results using the basic dataset were not promising. To address this, various dataset splits were tested using a K-fold Cross Validation method. This analysis revealed that certain darker or noisier images were significantly impacting the overall performance results.

Supposing that 61 images might not be enough for drawing solid conclusions, four types of data augmentation techniques were explored. All of these techniques led to improvements in performance metrics, with two techniques – translation and rotation – standing out by achieving a 98% F1-score.

In the context of health, a higher recall is generally more desirable than higher precision. Recall represents the rate of correctly identified real positives. A higher recall means fewer False Negatives (unhealthy patients) go undetected.

Between the two highlighted data augmentation techniques, rotation would be preferable, since the application of this technique resulted in a 100% recall, surpassing the 95% obtained with translation.

In the second phase, the identification of calcium presence in the aortic valve involved two distinct approaches: (1) a baseline heuristic method without Deep Learning (DL), and (2) a DL-based method using Neural Networks (NNs).

The heuristic approach employed a manual method to binarize images and classify them based on the presence of white pixels. As expected, this method produced low and inconsistent results, failing to provide solid conclusions.

Expecting better performance, a DL-based method was implemented, employing a NN with Transfer Learning from pre-trained models – as used for the object detector in the first phase. Initially, the original images, containing the entire heart, were used to train the model. However, the model's performance was around 50%. To enhance performance, images were subsequently cropped based on the original bounding boxes limits to focus on the region of interest containing the aortic valve.

Two pre-trained models were selected to be applied – EfficientNetB0 and MobileNetV3Small – because of their simpler architectures, designed for small datasets. When trained with the original dataset, the results were not promising. Trying to improve model's performance, two data treatment techniques were tested – balancing and data augmentation. Data augmentation consisted of rotations of the cropped images, and improved the results significantly, when using MobileNetV3Small pre-trained model. When combined with data balancing (making sure each class has a similar number of examples) the results got even higher, for both pre-trained models.

Since data balancing and augmentation resulted in promising results, this dataset was used to test different pre-trained models, for comparison. From six different pre-trained models applied, EfficientNetB0 and MobileNetV3Large resulted in the best performance, with 96% f1-score. These models are known for their simpler structure, which might have resulted in this case with a small dataset. MobileNetV3Large stood out by achieving 100% recall, and since high recall is important for identifying all cases of Aortic Stenosis and preventing undiagnosed cases, it was considered the best model in this context.

CHAPTER 5

Conclusion

5.1. Final Remarks

The main purpose of this dissertation was to create an automatic method for AS detection in echocardiography exams, through the identification of calcium structures in aortic valve, using DL applied to medical images. To achieve that, it was necessary to (1) create an object detector for aortic valve identification, and (2) developing an automatic classifier to identify the presence of calcium – a crucial indicator of AS.

Data selection was an essential process to obtain the necessary data for applying DL models and achieving the proposed goal. The data collection was conducted in compliance with confidentiality rules, in collaboration with Santa Maria's Hospital and Affidea clinics. Through anonymization, cleansing, and uniformization processes, a dataset containing 61 echocardiography images – all captured from the same heart view – was compiled for use in the modeling phase. Recognizing that this quantity of data might not be enough to train DL models and achieve satisfactory results, four data augmentation techniques – zooming, translation, rotation, and gamma contrast adjustments – were employed to generate additional images through transformations of the originals.

The final goal was to compile these two algorithms into an automated framework capable of recognizing the aortic valve in an echocardiogram and subsequently classifying the output, thereby indicating the presence of AS. Such an advancement in medical workflow would spare healthcare professionals from undertaking unnecessary, time-consuming tasks. Given the limited developments in this field, particularly in the context of echocardiography, as discerned from the conducted literature review, this would indeed represent an innovative advancement.

The object detection model, which aimed to automate the detection of the aortic valve, surpassed expectations, even with a relatively small dataset. The most successful model was built using Transfer Learning from the YOLOv5s network. To train this model, the originally collected dataset was subject to several data pre-processing procedures, which involved standardizing the images and implementing data augmentation to expand the dataset. This augmentation consisted of rotating all the images using two small angular variations each, resulting in a dataset three times the original size. This model demonstrated a 95% precision in detecting the aortic valve and achieved a 100% recall rate, meaning that all aortic valves were accurately located by the model in the echocardiography images.

In the subsequent phase, dedicated to developing an automated AS classifier, the results were not as satisfactory as those of the object detector. The most promising approach was based on DL, employing Transfer Learning from the EfficientNet model. The dataset was pre-processed to mitigate class imbalances, and rotated images were incorporated as an additional data augmentation method to train the model on an expanded dataset. This model demonstrated better consistency when compared with the others, achieving an accuracy of 63%, a precision of 62%, and a recall of 65%. Although this model only identified about 65% of the AS cases, it generated fewer instances of false alarms (where a healthy patient is misclassified as having AS) when compared with others. The remaining models demonstrated they were classifying the majority of the cases into the same class, which required a larger dataset for training to get around, helping the model to generalize the acquired knowledge to new images.

Throughout the process, the guidance of medical professionals played an essential role, ensuring that the results were aligned with practical objectives. This collaborative effort allowed for continuous adaptation of experiments, aiming to develop a viable solution that aligns with the proposed goal. As a result of these validation steps, it was determined that the Neural Network built for object detection performs a solution for aortic valve identification, improving medical procedures. In contrast, the obtained classifier requires further refinement to achieve higher levels of accuracy and establish itself as a reliable tool for diagnosing AS.

5.2. Future Work

Overall, the performance of the solutions obtained proved to be positive. However, there is room for further experimentation in the future, and continued development may yield better results. Both the object detection and classification solutions were built with specific parameters that can be adjusted to explore other variations, such as different network architectures or training settings.

The primary challenge encountered through this work was related to data. Acquiring the necessary data requires consistent communication with the clinic and special attention to compliance procedures, making it a time-intensive process. Subsequently, the dataset resulted in 61 images, which can be considered limited when compared to the typically larger datasets required to effectively train Neural Networks and achieve great results. Notably, the issue of limited data was apparently mitigated in the object detection model through the implementation of data augmentation, which did not yield the same level of improvement in the classification model.

To address this data quantity challenge, it is crucial to ensure that larger quantities of data are available for training these Neural Networks in future developments. Sufficient data enables the model to achieve a deeper understanding of the dataset and extract vital features. Consequently, the model becomes more adaptable to new data and has the potential to achieve higher performance levels.

Bibliography

[1] European Parliament, Directorate-General for Parliamentary Research Services, and P. Boucher, *Artificial intelligence – How does it work, why does it matter, and what we can do about it?* European Parliament, 2020. doi: 10.2861/44572.

[2] M. Roser, 'The brief history of artificial intelligence: The world has changed fast – what might be next?', Our World in Data. Accessed: Sep. 29, 2023. [Online]. Available: https://ourworldindata.org/brief-history-of-ai

[3] J.-S. Hong *et al.*, 'Automated coronary artery calcium scoring using nested U-Net and focal loss', *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1681–1690, 2022, doi: 10.1016/j.csbj.2022.03.025.

[4] G. A. Roth *et al.*, 'Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study', *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 2020, doi: 10.1016/j.jacc.2020.11.010.

[5] N. Gogin *et al.*, 'Automatic coronary artery calcium scoring from unenhanced-ECG-gated CT using deep learning', *Diagn. Interv. Imaging*, vol. 102, no. 11, pp. 683–690, 2021, doi: 10.1016/j.diii.2021.05.004.

[6] A. Faggiano, G. Santangelo, S. Carugo, G. Pressman, E. Picano, and P. Faggiano, 'Cardiovascular Calcification as a Marker of Increased Cardiovascular Risk and a Surrogate for Subclinical Atherosclerosis: Role of Echocardiography', *J. Clin. Med.*, vol. 10, no. 8, 2021, doi: 10.3390/jcm10081668.

[7] H. Baumgartner (chair) *et al.*, 'Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography', *Eur. Heart J. - Cardiovasc. Imaging*, vol. 18, no. 3, pp. 254–275, Mar. 2017, doi: 10.1093/ehjci/jew335.

[8] I. Wahlang *et al.*, 'Deep Learning Methods for Classification of Certain Abnormalities in Echocardiography', *Electronics*, vol. 10, no. 4, 2021, doi: 10.3390/electronics10040495.

[9] D. Bos and M. J. G. Leening, 'Leveraging the coronary calcium scan beyond the coronary calcium score', *Eur. Radiol.*, vol. 28, no. 7, pp. 3082–3087, Jul. 2018, doi: 10.1007/s00330-017-5264-3.

[10] M. J. Schuuring, I. Išgum, B. Cosyns, S. A. J. Chamuleau, and B. J. Bouma, 'Routine Echocardiography and Artificial Intelligence Solutions', *Front. Cardiovasc. Med.*, vol. 8, 2021, doi: 10.3389/fcvm.2021.648877.

[11] K. Seetharam, S. Shrestha, and P. P. Sengupta, 'Cardiovascular Imaging and Intervention Through the Lens of Artificial Intelligence', *Interv. Cardiol. 202116e31*, 2021, doi: 10.15420/icr.2020.04.

[12] R. Maragna *et al.*, 'Artificial Intelligence Based Multimodality Imaging: A New Frontier in Coronary Artery Disease Management', *Front. Cardiovasc. Med.*, vol. 8, Sep. 2021, doi: 10.3389/fcvm.2021.736223.

[13] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, 'A survey of the recent architectures of deep convolutional neural networks', *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[14] J. Chamberlin *et al.*, 'Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value', *BMC Med.*, vol. 19, no. 1, 2021, doi: 10.1186/s12916-021-01928-3.

[15] J. M. Wolterink, R. W. V. Hamersvelt, M. A. Viergever, T. Leiner, and I. Išgum, 'Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier', *Med. Image Anal.*, vol. 51, pp. 46–60, 2019, doi: 10.1016/j.media.2018.10.005.

[16] S. Lee *et al.*, 'Deep-learning-based coronary artery calcium detection from ct image', *Sensors*, vol. 21, no. 21, 2021, doi: 10.3390/s21217059.

[17] Y. Zhang, N. R. van der Werf, B. Jiang, R. van Hamersvelt, M. J. W. Greuter, and X. Xie, 'Motion-corrected coronary calcium scores by a convolutional neural network: a robotic simulating study', *Eur. Radiol.*, vol. 30, no. 2, pp. 1285–1294, 2020, doi: 10.1007/s00330-019-06447-7.

[18] F. N. Guilenea, M. E. Casciaro, A. F. Pascaner, G. Soulat, E. Mousseaux, and D. Craiem, 'Thoracic aorta calcium detection and quantification using convolutional neural networks in a large cohort of intermediate-risk patients', *Tomography*, vol. 7, no. 4, pp. 636–649, 2021, doi: 10.3390/tomography7040054.
[19] M. van Assen *et al.*, 'Automatic coronary calcium scoring in chest CT using a deep neural network in direct comparison with non-contrast cardiac CT: A validation study', *Eur. J. Radiol.*, vol. 134, 2021, doi: 10.1016/j.ejrad.2020.109428.
[20] 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', *J. Pers. Med.*, vol. 11, no. 7, 2021, doi: 10.3390/jpm11070598.
[21] C. Schröer, F. Kruse, and J. M. Gómez, 'A Systematic Literature Review on Applying CRISP-DM Process Model', *CENTERIS 2020 - Int. Conf. Enterp. Inf. Syst. ProjMAN 2020 - Int. Conf. Proj. Manag. HCist 2020 - Int. Conf. Health Soc. Care Inf. Syst. Technol. 2020 CENTERISProjMANHCist 2020*, vol. 181, pp. 526–534, Jan. 2021, doi: 10.1016/j.procs.2021.01.199.
[22] 'The-Crisp-DM-Methodology/ at main · almirgouvea/The-Crisp-DM-Methodology · GitHub'. Accessed: Dec. 27, 2023. [Online]. Available: https://github.com/almirgouvea/The-Crisp-DM-Methodology?search=1
[23] '#1 : CRISP DM Framework', Rucha Tare. Accessed: Dec. 22, 2023. [Online]. Available: https://ruchareads.wordpress.com/2021/03/29/1-crisp-dm-framework/
[24] M. J. Page *et al.*, 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ*, vol. 372, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
[25] L. B. van den Oever *et al.*, 'Application of artificial intelligence in cardiac CT: From basics to clinical practice', *Eur. J. Radiol.*, vol. 128, 2020, doi: 10.1016/j.ejrad.2020.108969.
[26] M. Klug *et al.*, 'A deep-learning method for the denoising of ultra-low dose chest CT in coronary artery calcium score evaluation', *Clin. Radiol.*, vol. 77, no. 7, pp. e509–e517, 2022, doi: 10.1016/j.crad.2022.03.005.
[27] W. Wang, L. Yang, S. Wang, Q. Wang, and L. Xu, 'An automated quantification method for the Agatston coronary artery calcium score on coronary computed tomography angiography', *Quant. Imaging Med. Surg.*, vol. 12, no. 3, pp. 1787–1799, 2022, doi: 10.21037/qims-21-775.
[28] D. H. Yang, 'Application of artificial intelligence to cardiovascular computed tomography', *Korean J. Radiol.*, vol. 22, pp. 1–12, 2021, doi: 10.3348/kjr.2020.1314.
[29] F. Yasmin *et al.*, 'Artificial intelligence in the diagnosis and detection of heart failure: the past, present, and future', *Rev. Cardiovasc. Med.*, vol. 22, no. 4, pp. 1095–1113, 2021, doi: 10.31083/j.rcm2204121.
[30] D. Eng *et al.*, 'Automated coronary calcium scoring using deep learning with multicenter external validation', *Npj Digit. Med.*, vol. 4, no. 1, 2021, doi: 10.1038/s41746-021-00460-1.
[31] P. M. Graffy, J. Liu, S. O'Connor, R. M. Summers, and P. J. Pickhardt, 'Automated segmentation and quantification of aortic calcification at abdominal CT: application of a deep learning-based algorithm to a longitudinal screening cohort', *Abdom. Radiol.*, vol. 44, no. 8, pp. 2921–2928, 2019, doi: 10.1007/s00261-019-02014-2.
[32] C. Xu *et al.*, 'Automatic coronary artery calcium scoring on routine chest computed tomography (CT): comparison of a deep learning algorithm and a dedicated calcium scoring CT', *Quant. Imaging Med. Surg.*, vol. 12, no. 5, pp. 2684–2695, 2022, doi: 10.21037/qims-21-1017.
[33] M. J. Emaus *et al.*, 'Bragatston study protocol: A multicentre cohort study on automated quantification of cardiovascular calcifications on radiotherapy planning CT scans for cardiovascular risk prediction in patients with breast cancer', *BMJ Open*, vol. 9, no. 7, 2019, doi: 10.1136/bmjopen-2018-028752.
[34] S. J. Al'Aref *et al.*, 'Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging', *Eur. Heart J.*, vol. 40, no. 24, pp. 1975–1986, 2019, doi: 10.1093/eurheartj/ehy404.
[35] W. Wang *et al.*, 'Coronary artery calcium score quantification using a deep-learning algorithm', *Clin. Radiol.*, vol. 75, no. 3, p. 237.e11-237.e16, Mar. 2020, doi: 10.1016/j.crad.2019.10.012.

[36] L. B. van den Oever *et al.*, 'Deep learning for automated exclusion of cardiac CT examinations negative for coronary artery calcium', *Eur. J. Radiol.*, vol. 129, 2020, doi: 10.1016/j.ejrad.2020.109114.
[37] S. G. M. van Velzen *et al.*, 'Deep learning for automatic calcium scoring in CT: Validation using multiple cardiac CT and chest CT protocols', *Radiology*, vol. 295, no. 1, pp. 66–79, 2020, doi: 10.1148/radiol.2020191621.
[38] A. Hoori, T. Hu, J. Lee, S. Al-Kindi, S. Rajagopalan, and D. L. Wilson, 'Deep learning segmentation and quantification method for assessing epicardial adipose tissue in CT calcium score scans', *Sci. Rep.*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-06351-z.
[39] B. Jiang, N. Guo, Y. Ge, L. Zhang, M. Oudkerk, and X. Xie, 'Development and application of artificial intelligence in cardiac imaging', *Br. J. Radiol.*, vol. 93, no. 1113, 2020, doi: 10.1259/bjr.20190812.
[40] B. D. de Vos, J. M. Wolterink, T. Leiner, P. A. de Jong, N. Lessmann, and I. Isgum, 'Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT', *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2127–2138, 2019, doi: 10.1109/TMI.2019.2899534.
[41] G. Singh *et al.*, 'End-to-end, pixel-wise vessel-specific coronary and aortic calcium detection and scoring using deep learning', *Diagnostics*, vol. 11, no. 2, 2021, doi: 10.3390/diagnostics11020215.
[42] M. T. Winkelmann *et al.*, 'Fully Automated Artery-Specific Calcium Scoring Based on Machine Learning in Low-Dose Computed Tomography Screening', *RoFo Fortschritte Auf Dem Geb. Rontgenstrahlen Bildgeb. Verfahr.*, vol. 194, no. 7, pp. 763–770, 2022, doi: 10.1055/a-1717-2703.
[43] J. Lee *et al.*, 'Fully Automatic Coronary Calcium Score Software Empowered by Artificial Intelligence Technology Validation Study Using Three CT Cohorts', *KOREAN J. Radiol.*, vol. 22, no. 11, pp. 1764–1776, Nov. 2021, doi: 10.3348/kjr.2021.0148.
[44] P. T. Lauzier, R. Avram, D. Dey, P. Slomka, J. Afilalo, and B. J. W. Chow, 'The Evolving Role of Artificial Intelligence in Cardiac Image Analysis', *Can. J. Cardiol.*, vol. 38, no. 2, pp. 214–224, 2022, doi: 10.1016/j.cjca.2021.09.030.
[45] J. Waltz, M. Kocher, J. Kahn, M. Dirr, and J. Burt, 'The Future of Concurrent Automated Coronary Artery Calcium Scoring on Screening Low-Dose Computed Tomography', *CUREUS*, vol. 12, no. 6, Jun. 2020, doi: 10.7759/cureus.8574.
[46] C. Antoniades, F. W. Asselbergs, and P. Vardas, 'The year in cardiovascular medicine 2020: Digital health and innovation', *Eur. Heart J.*, vol. 42, no. 7, pp. 732–739, 2021, doi: 10.1093/eurheartj/ehaa1065.
[47] L. Hahn, K. Baeumler, and A. Hsiao, 'Artificial intelligence and machine learning in aortic disease', *Curr. Opin. Cardiol.*, vol. 36, no. 6, pp. 695–703, Nov. 2021, doi: 10.1097/HCO.0000000000000903.
[48] F. Yang *et al.*, 'Automated Analysis of Doppler Echocardiographic Videos as a Screening Tool for Valvular Heart Diseases', *JACC Cardiovasc. Imaging*, vol. 15, no. 4, pp. 551–563, Apr. 2022, doi: 10.1016/j.jcmg.2021.08.015.
[49] Y. Takeshita *et al.*, 'Coronary artery calcium score may be a novel predictor of COVID-19 prognosis: A retrospective study', *BMJ Open Respir. Res.*, vol. 8, no. 1, 2021, doi: 10.1136/bmjresp-2021-000923.
[50] D. J. Winkel *et al.*, 'Deep learning for vessel-specific coronary artery calcium scoring: validation on a multi-centre dataset', *Eur. Heart J. Cardiovasc. Imaging*, vol. 23, no. 6, pp. 846–854, 2022, doi: 10.1093/ehjci/jeab119.
[51] N. Zhang *et al.*, 'Fully automatic framework for comprehensive coronary artery calcium scores analysis on non-contrast cardiac-gated CT scan: Total and vessel-specific quantifications', *Eur. J. Radiol.*, vol. 134, 2021, doi: 10.1016/j.ejrad.2020.109420.
[52] R. Gal *et al.*, 'Identification of Risk of Cardiovascular Disease by Automatic Quantification of Coronary Artery Calcifications on Radiotherapy Planning CT Scans in Patients with Breast Cancer', *JAMA Oncol.*, vol. 7, no. 7, pp. 1024–1032, 2021, doi: 10.1001/jamaoncol.2021.1144.

[53] J. Lee *et al.*, 'Prediction of Coronary Artery Calcium Score Using Machine Learning in a Healthy Population', *J. Pers. Med.*, vol. 10, no. 3, Sep. 2020, doi: 10.3390/jpm10030096.
[54] 'DICOM', DICOM. Accessed: Oct. 26, 2023. [Online]. Available: https://www.dicomstandard.org
[55] 'Home', OpenCV. Accessed: Oct. 05, 2023. [Online]. Available: https://opencv.org/
[56] M. Hussain, 'YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection', *Machines*, vol. 11, no. 7, 2023, doi: 10.3390/machines11070677.
[57] 'MAXQDA | All-In-One Qualitative & Mixed Methods Data Analysis Tool', MAXQDA. Accessed: Oct. 26, 2023. [Online]. Available: https://www.maxqda.com/
[58] C. Lee, 'How many images do you need for object detection?', Medium. Accessed: Oct. 01, 2023. [Online]. Available: https://changsin.medium.com/how-many-images-do-you-need-for-object-detection-d33185629843
[59] 'What is data augmentation in deep learning?' Accessed: Oct. 20, 2023. [Online]. Available: https://www.calipsa.io/blog/what-is-data-augmentation-in-deep-learning
[60] 'imgaug — imgaug 0.4.0 documentation'. Accessed: Oct. 17, 2023. [Online]. Available: https://imgaug.readthedocs.io/en/latest/
[61] 'What Is Object Detection?' Accessed: Oct. 01, 2023. [Online]. Available: https://www.mathworks.com/discovery/object-detection.html
[62] 'TensorFlow'. Accessed: Oct. 05, 2023. [Online]. Available: https://www.tensorflow.org/?hl=pt-br
[63] 'Keras: Deep Learning for humans'. Accessed: Oct. 05, 2023. [Online]. Available: https://keras.io/
[64] 'PyTorch'. Accessed: Oct. 05, 2023. [Online]. Available: https://www.pytorch.org
[65] H. Kim, 'Classification Performance Metric with Python Sklearn', Analytics Vidhya. Accessed: Oct. 07, 2023. [Online]. Available: https://medium.com/analytics-vidhya/classification-performance-metric-with-python-sklearn-d8342ac25898
[66] K. E. Koech, 'Confusion Matrix and Object Detection', Medium. Accessed: Oct. 08, 2023. [Online]. Available: https://towardsdatascience.com/confusion-matrix-and-object-detection-f0cbcb634157
[67] R. Patro, 'Cross Validation: K Fold vs Monte Carlo', Medium. Accessed: Oct. 22, 2023. [Online]. Available: https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b
[68] K. Team, 'Keras documentation: EfficientNet B0 to B7'. Accessed: Oct. 05, 2023. [Online]. Available: https://keras.io/api/applications/efficientnet/
[69] K. Team, 'Keras documentation: MobileNet, MobileNetV2, and MobileNetV3'. Accessed: Oct. 05, 2023. [Online]. Available: https://keras.io/api/applications/mobilenet/
[70] K. Team, 'Keras documentation: ResNet and ResNetV2'. Accessed: Oct. 05, 2023. [Online]. Available: https://keras.io/api/applications/resnet/
[71] R. Vidiyala, 'Performance Metrics for Classification Machine Learning Problems', Medium. Accessed: Oct. 09, 2023. [Online]. Available: https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007
[72] C.-H. Chiang, C.-L. Weng, and H.-W. Chiu, 'Automatic classification of medical image modality and anatomical location using convolutional neural network', *PLOS ONE*, vol. 16, no. 6, p. e0253205, Jun. 2021, doi: 10.1371/journal.pone.0253205.
[73] B. K. Mishra, D. Thakker, S. Mazumdar, D. Neagu, M. Gheorghe, and S. Simpson, 'A novel application of deep learning with image cropping: a smart city use case for flood monitoring', *J. Reliab. Intell. Environ.*, vol. 6, no. 1, pp. 51–61, Mar. 2020, doi: 10.1007/s40860-020-00099-x.
[74] 'Introduction to Balanced and Imbalanced Datasets in Machine Learning'. Accessed: Oct. 05, 2023. [Online]. Available: https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/
[75] 'tf.keras.applications.efficientnet.EfficientNetB0 | TensorFlow v2.14.0', TensorFlow. Accessed: Nov. 22, 2023. [Online]. Available:

https://www.tensorflow.org/api_docs/python/tf/keras/applications/efficientnet/EfficientNetB0
[76] ‘tf.keras.applications.MobileNetV3Small | TensorFlow v2.14.0’, TensorFlow. Accessed: Nov. 22, 2023. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/applications/MobileNetV3Small
[77] ‘Dealing with very small datasets’. Accessed: Nov. 22, 2023. [Online]. Available: https://kaggle.com/code/rafjaa/dealing-with-very-small-datasets