UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

João Francisco Barreto da Silva Martins

Diagnóstico Automático de Cardiopatia Reumática em Exames Ecocardiográficos

Belo Horizonte
2021

João Francisco Barreto da Silva Martins

Diagnóstico Automático de Cardiopatia Reumática em Exames Ecocardiográficos

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientadora: Gisele Lobo Pappa
Coorientador: Erickson Rangel do Nascimento

Belo Horizonte
2021

João Francisco Barreto da Silva Martins

Automatic Diagnosis of Rheumatic Heart Disease in Echocardiographic Exams

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Gisele Lobo Pappa
Co-Advisor: Erickson Rangel do Nascimento

Belo Horizonte
2021

Martins, João Francisco Barreto da Silva.

M386a Automatic diagnosis of rheumatic heart disease in echocardiographic exams [manuscrito] / João Francisco Barreto da Silva Martins. — Belo Horizonte, 2021.
84 f. il.; 29 cm.

Orientadora: Gisele Lobo Pappa.
Coorientador: Erickson Rangel do Nascimento
Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Referências: f.74-83.

1. Computação – Teses. 2. Visão por computador – Teses. 3. Aprendizagem de máquina – Teses. 4. Aprendizado profundo – Teses. 5. Ecocardiografia – Teses. 6. Cardiopatia reumática – Teses. I. Pappa, Gisele Lobo. II. Nascimento, Erickson Rangel do III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV. Título.

CDU 519.6*82 (043)

Ficha Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Automatic Diagnosis of Rheumatic Heart Disease in Echocardiographic Exams

JOÃO FRANCISCO BARRETO DA SILVA MARTINS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. GISELE LOBO PAPPA - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ERICKSON RANGEL DO NASCIMENTO - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ROBERTO DE ALENCAR LOTUFO
Faculdade de Engenharia Elétrica e de Computação - UNICAMP

PROF. BRUNO RAMOS NASCIMENTO
Departamento de Clínica Médica - Universidade Federal de Minas Gerais

Belo Horizonte, 13 de Dezembro de 2021.

To my family and friends, which enduring support was vital during this journey

Acknowledgments

First and foremost, I would like to thank my parents, Andrea and Antônio, and my sisters, Giovanna and Virgínia, for always supporting me, showing unconditional love, instigating my curiosity, and motivating me to go forward even when things seemed grim.

This accomplishment would not be possible without the work and brilliant guidance given by my advisors, Gisele Pappa and Erickson Nascimento. You have gifted me with knowledge that goes beyond the technical information needed to finish this dissertation and which I will take with me forever. Also, my sincere appreciation for Bruno Nascimento, our super collaborator and the medical mastermind behind this research.

Thanking WiRED International is important, as their exceptional content production served as a foundation stone in my comprehension of Echocardiography and Rheumatic Heart Disease, in addition to having given me permission to use their images in this document.

I would also like to thank my colleagues and friends at Hekima, for all the productive discussions about artificial intelligence, science, and philosophy. Special thanks to Bruno Laporais, Daniel Galinkin, Daniel Vieira, Fernando Bombardelli, Luis Rios, Matheus Caldas, Murilo Menezes, Natércia Aguilar, Raphael Ottoni, Ronald Pereira, Thiago Cardoso, and Vitor Oliveira.

Finally, but essential, I want to express my deep gratitude for my collaborator and great friend, Edson Araujo. Your bright ideas, hour-long discussions, and encouraging words were the remaining pieces needed for me to conclude this work.

"*Observar e absorver*"
(Eduardo Marinho)

Resumo

Cardiopatia Reumática (CR) afeta aproximadamente 39 milhões de pessoas ao redor do mundo e é a doença cardíaca adquirida mais comum entre crianças e adolescentes. A doença é responsável por mais de 300.000 mortes anualmente e figura entre as principais causas de morte e invalidez em países de baixa e média renda mas pode ser evitada se detectada precocemente. Ecocardiogramas são o padrão ouro para o diagnóstico de CR, sendo uma ferramenta muito eficaz para sua identificação enquanto latente. Devido ao custo de equipamento e à escassez de mão de obra qualificada, a adoção em massa de programas de rastreamento para detecção precoce e prevenção da progressão da doença em áreas endêmicas ainda é severamente restrita. Avanços tecnológicos recentes diminuíram o custo de máquinas ecocardiográficas portáteis, porém a lacuna de mão de obra qualificada permanece e poderia ser preenchida através da implementação de aplicações para diagnóstico auxiliado por computador.

Neste trabalho, abordamos os desafios do diagnóstico automático de CR em exames ecocardiográficos convencionais. Não há literatura prévia sobre o assunto, e hipotetizamos que os métodos desenvolvidos para tarefas relacionadas provavelmente não funcionariam tão bem devido à negligência de informações temporais. Para testar essa hipótese, comparamos o desempenho de uma rede neural convolucional (RNC) 3D com um modelo de tamanho semelhante da literatura ao prever a presença de CR em cada vídeo. Também propomos uma estratégia de agregação mais sofisticada para emitir o diagnóstico de um exame completo, que é supervisionada e baseado nos momentos da distribuição de confiança para as previsões de vídeo do classificador anterior. Experimentos mostram que o modelo com noção temporal e a estratégia de agregação supervisionada são significativamente melhores na tarefa de diagnóstico de CR.

Finalmente, apresentamos uma rede neural convolucional de dois fluxos em uma configuração de aprendizado multitarefa que usa uma RNC 2D como extrator de características, mas que ainda assim é capaz de incorporar informações temporais na previsão por meio de mecanismos de atenção. Além disso, propomos uma estratégia de agregação não supervisionada que é centrada na detecção de vídeos fora da distribuição como instâncias ruidosas, eventualmente removendo-os do processo de diagnóstico final. Ao levar em conta rótulos de anormalidades funcionais do coração como tarefas auxiliares durante o treino, nosso novo método não só é capaz de superar significativamente outros métodos tomados linhas de base com uma acurácia de 71,18% mas também é capaz de fornecer informações consistentes sobre seu processo de tomada de decisão em múltiplos

níveis, principalmente como visualizações temporais (quadros relevantes no vídeo) e espaciais (estruturas relevantes em um quadro). Direções para a adoção dessa tecnologia no mundo real são discutidas.

Palavras-chave: aprendizado de máquina, aprendizado profundo, visão computacional, ecocardiografia, cardiopatia reumática

Abstract

Rheumatic Heart Disease (RHD) affects an estimated 39 million people worldwide and is the most common acquired heart disease in children and young adults. The disease is responsible for more than 300,000 deaths annually and ranks as a leading cause of death and disability in low- and middle-income countries but is preventable if detected early. Echocardiograms are the gold standard for diagnosis of RHD, being a very effective tool for its identification while latent. Due to equipment costs and a shortage of skilled experts, the adoption of widespread screenings for early detection and prevention of disease progression in endemic areas is still severely restricted. Recent technological advancements increased the affordability of portable echocardiographic machines, but the gap of expert shortage remains and could be bridged by the implementation of computer-aided diagnosis applications.

In this work, we address the challenges of automatic diagnosis of RHD in conventional echocardiographic exams. There is no previous literature on the subject, and we hypothesized that methods developed for related tasks were unlikely to work well due to their disregard for temporal information. To test this hypothesis, we compare the performance of a 3D convolutional neural network (CNN) with a similar-sized model from the literature when predicting the presence of RHD in each video. We also propose a more sophisticated aggregation strategy to issue a whole exam diagnosis, which is supervised and based on the moments of the confidence distribution for the video predictions of the previous classifier. Experiments show that the temporal-aware model and the supervised aggregation strategy are significantly better at the task of RHD diagnosis.

Finally, we present a two-stream convolutional neural network in a multi-task learning setup that uses a 2D CNN as a feature extractor but can nonetheless incorporate temporal information in the prediction through attention mechanisms. Furthermore, we propose an unsupervised aggregation strategy centered around detecting out-of-distribution videos as noisy instances, ultimately removing them from the final diagnosis process. By leveraging labels of functional abnormalities of the heart as auxiliary tasks during training, our new method is not only able to significantly outperform other baselines with an accuracy of 71.18% but is also able to provide consistent information about its decision-making process in multiple levels, mainly as temporal (relevant frames in the video) and spatial (relevant structures in a frame) visualizations. Directions for real-world adoption of this technology are discussed.

Keywords: machine learning, deep learning, computer vision, echocardiography, rheumatic heart disease

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Cardiovascular disease (CVD) is the leading cause of mortality worldwide, with an estimated number of deaths of 18.6 million individuals per year [94]. Even though CVDs are considered an expanding threat to global health, socioeconomic, racial, and ethnic differences still play a crucial role in access to cardiovascular care [28, 26]. Rheumatic Heart Disease (RHD) — damaged heart valves derived from acute rheumatic fever (ARF) — affects an estimated 39 million people worldwide [56] and is the most common acquired heart disease in children and young adults [74, 80]. As a neglected disease [66], RHD ranks as a leading cause of death and disability in low- and middle-income countries (LMICs) but can be treated if detected in its early stages. Secondary prophylaxis in the form of regular penicillin injections can be initiated to prevent new episodes of ARF, avoiding further valve damage and progression of the disease [103]. In 2013, the Brazilian Public Health System reported $5,169$ hospitalizations related to ARF, and $8,841$ related to chronic RHD, at a cost of 33 million USD, mostly related to cardiovascular surgeries [89].

Thanks to the recent technological advances, echocardiography is more cost-effective and widely available [33]. Echocardiography is crucial for diagnosing a range of heart conditions [33] and reducing CVD-related deaths [33, 88, 81, 126]. In particular, echocardiograms have emerged as the gold standard for RHD diagnosis [32] and as a very effective tool for early detection of latent RHD, identifying 10 times more subclinical disease cases when compared with auscultation [70, 90]. Following guidelines published by the World Heart Federation in 2012 [88], an experienced echocardiographer can leverage findings related to structural (morphological) and functional abnormalities in the mitral valve and aortic valve to issue a diagnosis for RHD. However, the availability of skilled professionals has proven to be insufficient in underdeveloped regions, creating a gap in which computer-aided diagnosis of cardiac images can potentially help fill. Additional strategies to overcome the shortage of experts include task-shifting of imaging acquisition to non-physicians and utilization of telemedicine for remote diagnosis. Such approaches are made even more practical in resource-poor settings by the utilization of ultraportable handheld ultrasound devices.

The application of artificial neural networks to conventional 2-dimensional echocardiographic data dates back to 1990 [24]. In recent years the number of publications in the

field has risen considerably due to the popularization of deep learning (DL) [71]. Many medical fields, such as oncology and pneumology, have also seen successful applications of DL methods for disease detection [53, 112, 86, 98].

Concerning conventional echocardiograms, DL literature mainly comprises studies on echocardiogram viewpoint (view) identification [42, 67, 68, 125], heart chamber segmentation [125, 21], and classification of heart disease [65, 68, 44, 125], primarily applied for morphological rather than functional abnormalities. None of the disease-related research directly addresses valve abnormalities, and virtually all of the research uses a frame by frame (2D) approach to process images, discarding the temporal relation encoded in video clips.

Echocardiography identification of RHD, especially the subtle findings of subclinical disease, is highly dependent on the behavior of cardiac structures and blood flow across sequences of frames in a video. Therefore, it is unlikely that an approach that disregards temporal information would achieve the best performance.

1.1 Objectives and Contributions

This dissertation presents the first proposals to address the challenges of automatic RHD diagnosis in conventional echocardiographic exams. We aim to identify the traits that differentiate the task at hand from related works in echocardiology and propose machine learning methods that solve this problem even in data-poor settings.

The main goal of this work is to present a first step towards creating effective and interpretable methods for automatic diagnosis of RHD, which can reduce the cost of RHD screenings, increasing their coverage to possibly decrease the still heavy burden of RHD in endemic regions. The implementation of these methods in the real world can be done through a cloud-application for telemedicine diagnosis or by embedding them in screening devices for direct utilization at the point-of-care during screening programs.

The main contributions of this work are summarized as:

- A framework for using machine learning in the automated diagnosis of cardiac diseases that depends on temporal information and aggregation of multiple video predictions into a final diagnosis for a single exam while analyzing the decision-making process through multiple layers of interpretability.

- Two deep neural network architectures for RHD diagnosis in videos:

  - A 3D CNN based architecture as a baseline that is the currently published state-of-the-art [72];
  - A new attention-based multi-task architecture able to train in a regime of small datasets and provide interpretability in multiple levels.

- Two strategies for aggregating individual video predictions into a final exam diagnosis for RHD:
  - A supervised aggregation strategy based in meta-learning of the distributions of video predictions;
  - An unsupervised aggregation strategy based on a sparse regularization formulation that tackles out-of-distribution samples and provides another layer of interpretability.

1.2 Publications and Awards

In the following, we list the published works that are a direct contribution of this dissertation:

- B. R. Nascimento, **J. F. Martins**, E. R. Nascimento, G. L. Pappa, *et. al.* Deep Learning for Automatic Identification of Rheumatic Heart Disease in Echocardiographic Screening Images: Data from the PROVAR-ATMOSPHERE Study. In *Journal of the American College of Cardiology (JACC)*, 2020. Extended Abstract.
- B. R. Nascimento, **J. F. Martins**, E. R. Nascimento, G. L. Pappa, *et. al.* Spatial-temporal Deep Learning for Automatic Identification of Rheumatic Heart Disease in Echocardiographic Screening Images - Data from the PROVAR-ATMOSPHERE Study. In *Journal of the American College of Cardiology (JACC)*, 2021. Extended Abstract.
- **J. F. Martins**, E. R. Nascimento, B. R. Nascimento, C. A. Sable, *et. al.* Towards Automatic Diagnosis of Rheumatic Heart Disease on Echocardiographic Exams Through Video-based Deep Learning. In *Journal of the American Medical Informatics Association (JAMIA)*, 2021. Full Paper.

Also, the following awards have been received regarding this dissertation:

- Best Work (Main Track) — *75th Brazilian Congress of Cardiology* (2020). Work described in Chapter 3.

- Best Work (Main Track) — *30th Minas Gerais' Congress of Cardiology* (2021). Work described in Chapter 4.

1.3 Outline

The remainder of this document is organized as follows:

- **Chapter 2** discusses the clinical and computational backgrounds for this dissertation while also addressing related works in computer vision applied to echocardiography.
- **Chapter 3** proposes and evaluates a strong baseline method for the diagnosis of RHD. It comprises a 3D CNN for video classification and a supervised meta-classifier that aggregates video predictions into the final diagnosis.
- **Chapter 4** proposes a two-stream attention-based 2D CNN within a multi-task learning setup and an unsupervised sparse voting strategy for exam diagnosis. The method is compared to the baselines, and its interpretability capabilities are discussed.
- **Chapter 5** concludes this dissertation by reviewing our main contributions and proposing future research directions.

Chapter 2

Background

In this chapter, the clinical and computational concepts relevant to this work are presented in more detail. Finally, works related to automatic echocardiogram analysis with computer vision methods are presented and discussed in the context of this dissertation.

2.1 Clinical Background

Permission to use images in this section has been granted in all cases. Images sourced from [13] are available under a Creative Commons Attribution License 4.0. Images sourced from [36] were created by co-author Caroline Watson, and WiRED International has explicitly given permission to use.

2.1.1 Heart

Anatomy and Morphology

The heart is the organ responsible for the circulation of blood through the body of most animals. In humans, it is located medially between the lungs within the thoracic cavity and separated from other structures in the mediastinum by a tough membrane known as the pericardium. Figure 2.1 shows the heart's position and also identifies some of its internal and surrounding structures.

The heart is composed of 4 chambers, with each side being made up of an atrium and a ventricle. An atrium receives blood from external structures and contracts to push it to the ventricle that follows. Ventricles, on the other hand, function by pumping the blood out of the organ. Figure 2.2 depicts the heart's internal structures and the major blood vessels to which they are connected.

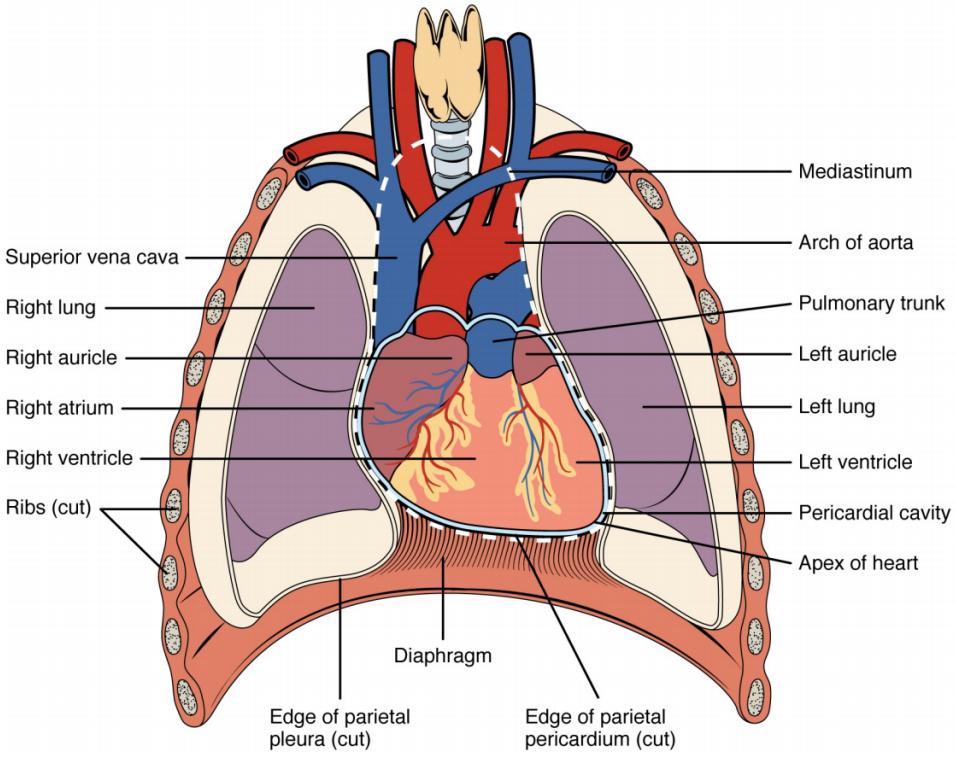


Figure 2.1: **Position of the heart in the thorax.** Image adapted from [13].

Present at the end of each chamber in the heart, there is a valve, made of 2 or 3 leaflets, which prevents blood that flows through it from returning to the previous chamber. Therefore, in a heart with normal function, blood flow is expected to be unidirectional. The valve between the right atrium and the right ventricle is called the tricuspid valve, while the one found where the right ventricle touches the pulmonary trunk is named the pulmonary valve. Between the atrium and the ventricle in the left side of the heart there is the mitral (bicuspid) valve, with the aortic valve located where the left ventricle connects to the base of the aorta.

The valves between the atria and ventricles, generically called atrioventricular valves, have an essential mechanism to ensure they function correctly. Chord-like tendons, known as chordae tendinae, connect their leaflets to the papillary muscles, which are anchored to ventricular walls and prevent the valves from inverting into the atria. Valves at the end of ventricles, generically called semilunar valves, have no chordae tendinae or papillary muscles associated to them.

Finally, the heart wall is composed of three different layers: epicardium, myocardium, and endocardium. The epicardium, which is also the innermost layer of the pericardium, is made of connective tissue and protects the heart by lubricating it to prevent friction during cardiac activity. The myocardium is where the cardiac muscle is located, which contracts through electrical stimulation to enable to heart to work as a blood pump. The inside of the heart is lined with the endocardium, which also compose valve leaflets along with additional connective tissue [13, 36].

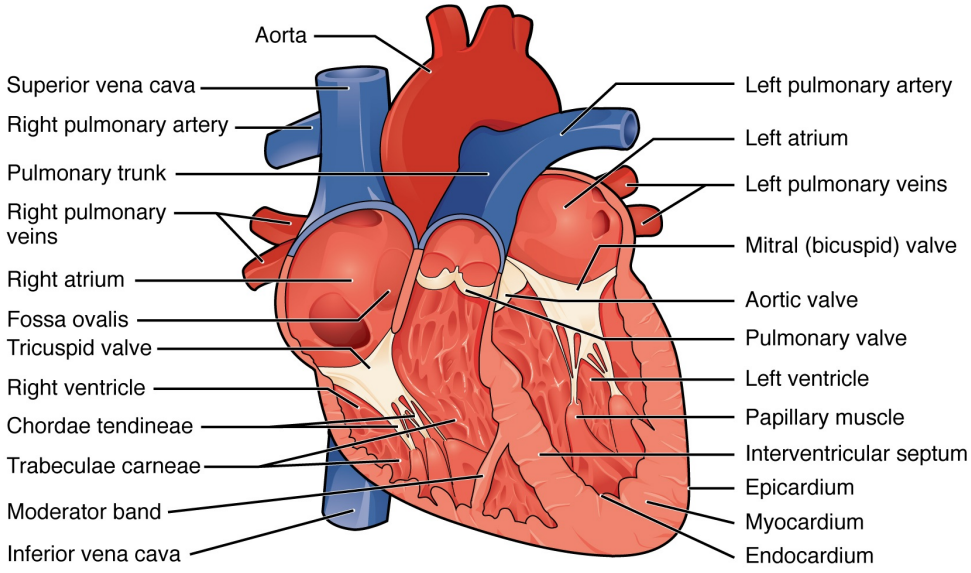


Figure 2.2: **Internal heart structures and external vessels connected to them.** Image adapted from [13].

Physiology

The heart's primary function is to take deoxygenated blood to the lungs and then pump the oxygen-rich blood that comes out of them to the rest of the body. The process of pumping blood through the 4 chambers of the heart is called a cardiac cycle, which occur from the beginning of a heartbeat to the beginning of the next.

The cardiac cycle can be divided into two basic phases, which both atria and ventricles undergo: systole (contraction) and diastole (relaxation). During ventricular systole, ventricles simultaneously contract, forcing blood out of the open semilunar valves. At the same time, atrial diastole is happening, with both relaxed atria being filled with blood, while the atrioventricular valves remain closed. During ventricular diastole and atrial systole the opposite is true: semilunar valves close while atrioventricular valves open letting the blood inside the contracting atria fill the relaxed ventricles. For clarity purposes, from now on, we will use the terms systole and diastole, always referring to their ventricular instances, as their atrial counterparts can be inferred from that.

Figure 2.3 depicts the circulation of blood through the heart during a cardiac cycle. Deoxygenated blood with high amounts of carbon dioxide comes from the superior vena cava into the right atrium. Afterward, it flows to the right ventricle and then into the lungs through the pulmonary arteries. Gas exchange happens in the pulmonary capillaries to remove carbon dioxide and enrich the blood with oxygen. Subsequently, the blood returns to the heart via the pulmonary veins connected to the left atrium. It is then pushed to the left ventricle, which finally pumps the blood back into the body. This blood will eventually return to the heart with low amounts of oxygen after exchanging gases in systemic capillaries [13, 36].

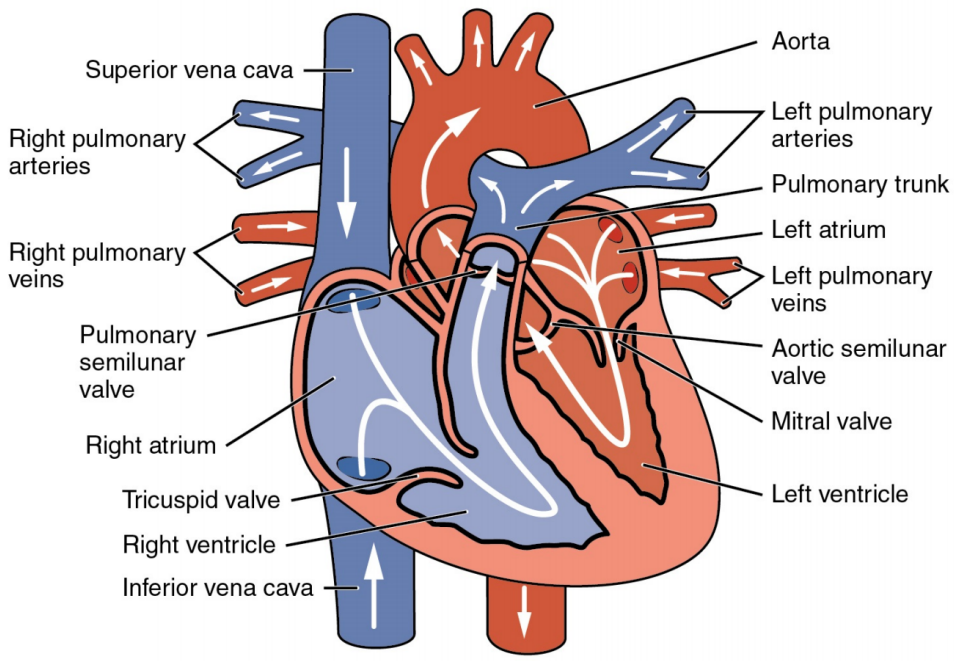


Figure 2.3: **Blood circulation through the heart's internal structures and external vessels.** Image adapted from [13].

2.1.2 Rheumatic Heart Disease

Burden, Symptoms and Causes

Rheumatic heart disease (RHD) is the leading acquired heart disease in children and young adults in the world [74, 80], affects 39 million people worldwide [56], and is responsible for more than 300,000 deaths annually according to the most recent estimate [75]. It is associated with household crowding and poverty [25], and ranks as a leading cause of death and disability in low- and middle-income countries (LMICs) [103] despite having almost completely disappeared from wealthy countries [116].

RHD is a chronic disease that results from episodes of acute rheumatic fever (ARF). ARF, in turn, is a complication of Group A Streptococcal (GAS) Pharyngitis, which is informally called strep throat. Figure 2.4 depicts the progression from GAS Pharyngitis to RHD. Strep throat occurrences are common, and in most people, usually resolve without treatment. However, there is a small risk of developing an autoimmune response which results in ARF about 3 weeks later [32]. The most common clinical symptoms of ARF are: large joint pain or swelling, acute fever, choreiform movements, and heart inflammation (carditis) [18]. Recurrent episodes of ARF, or a single very severe one, can result in permanent damage to valves in the heart due to carditis, which is diagnosed as RHD.

Valves on the left side of the heart (mitral and aortic) are predominantly affected by RHD, with damage to the mitral valve being the most common [18]. Damage to the

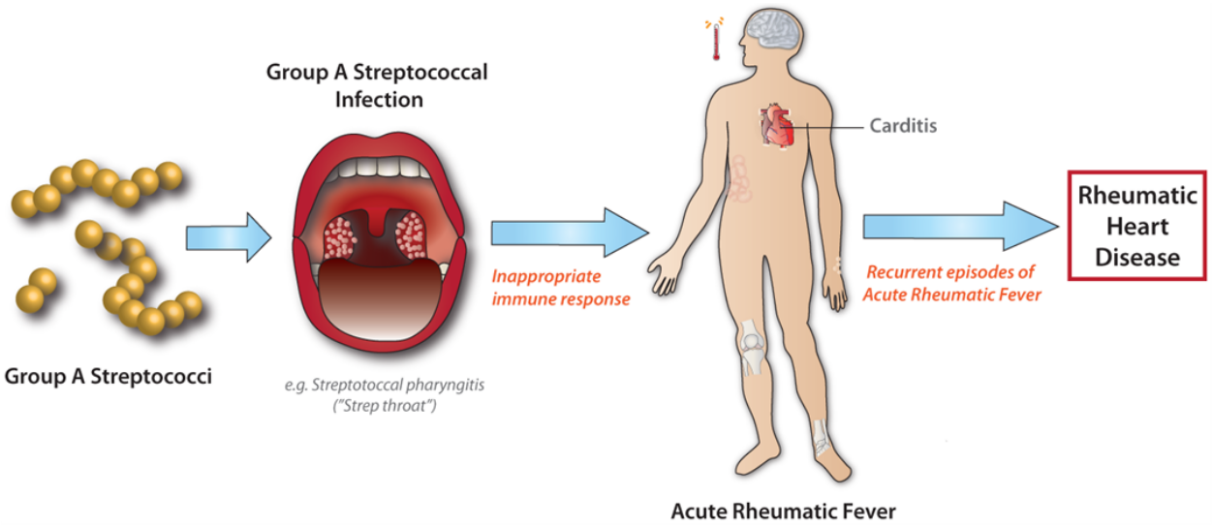


Figure 2.4: **Progression from Group A Streptococcal Pharyngitis to Rheumatic Heart Disease.** Image reproduced from [36].

valves results in morphological abnormalities such as scarring, thickness, and stiffness of the tissue, compromising their function. They may also begin not to close properly (regurgitation) or open fully (stenosis), resulting in leakage of blood back through the valve or poor blood flow to the next chamber, respectively. Figure 2.4 visually depicts valve regurgitation and stenosis. Stenosis on both valves tends to occur later in the disease, often co-existing with regurgitation [36].

If heart valves are not working as they should, the cardiovascular system strains itself to maintain proper circulation around the body. The heart can cope with this temporarily, but heart failure can develop with time, leading to worse complications and even death. Heart failure affects around 50% of patients with RHD, being the most common consequence of the disease, but atrial fibrillation, infective endocarditis, and even stroke are possible outcomes [32, 36].

Diagnosis

The diagnosis of RHD can be done clinically through the assessment of the patient's ARF history, combined with the presence of pathological murmur, exercise-induced chest pain, shortness of breath, heart failure, syncope, palpitations, atrial fibrillation, or stroke [32]. However, a long latent phase of asymptomatic valvular heart disease, often without any preceding history or symptoms of ARF, is the most common scenario [60].

Much subtler signals of RHD can be observed through cardiac imaging techniques, which allow observations of morphological changes to the valves and lone blood jets opposing the expected flow between steps of the cardiac cycle. Echocardiography is the current gold standard for RHD diagnosis [32] being a very effective tool for the detection of latent RHD, identifying 10 times more subclinical disease cases when compared with auscultation with a stethoscope [70, 90].

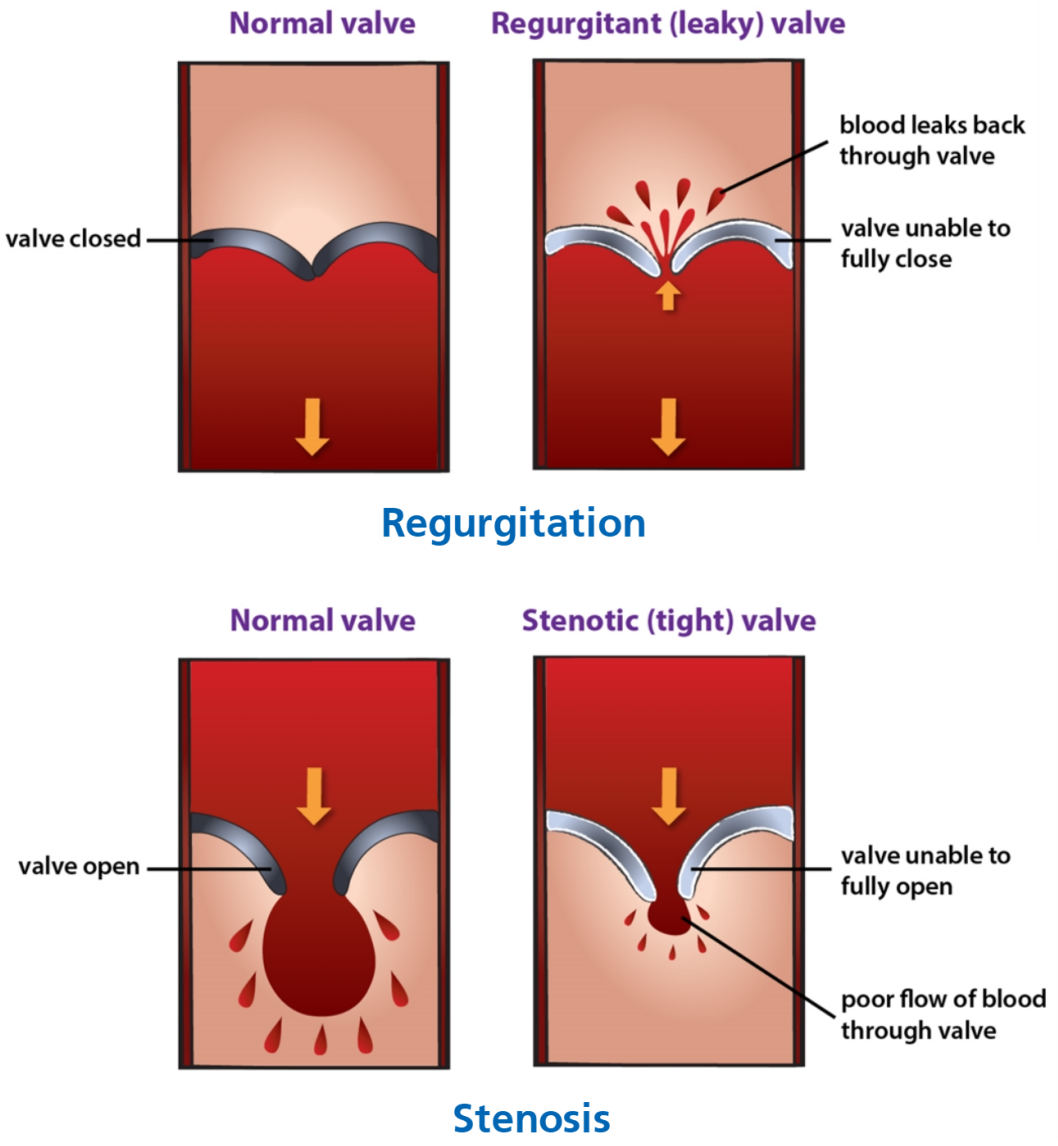


Figure 2.5: **Heart valve regurgitation and stenosis.** Image adapted from [36].

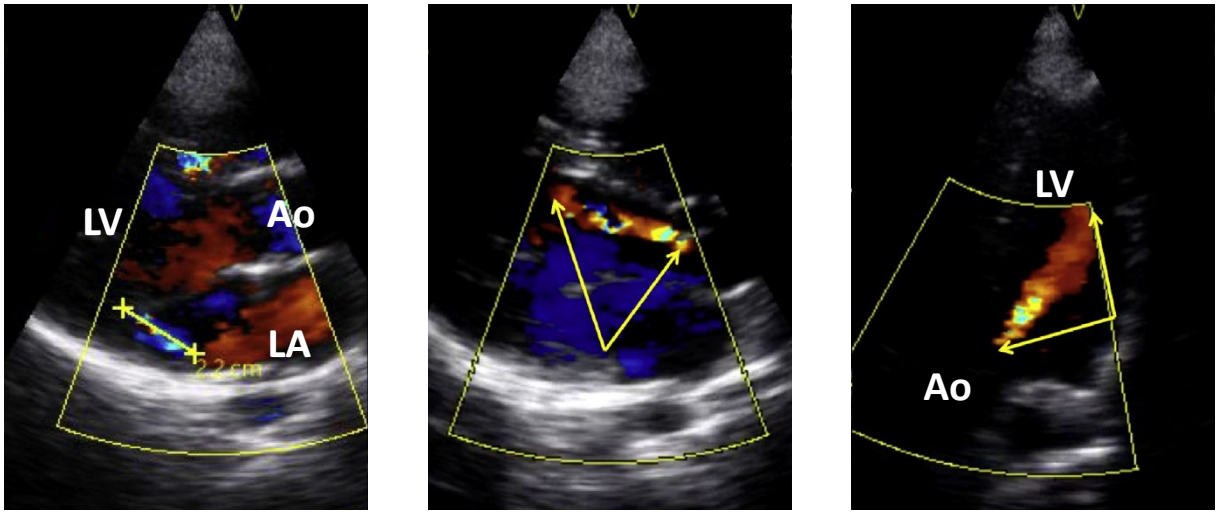
In 2012, the World Heart Federation (WHF) published the first evidence-based echocardiographic diagnosis guidelines [88], which, along with recent advances in echocardiographic technology that reduced acquisition cost and improved portability, made this the preferred technology for RHD diagnosis [88, 90, 10]. With these guidelines at hand, it is possible to issue a diagnosis for RHD by assessing echocardiographic findings: the combination of pathological left-sided regurgitation/stenosis and morphological changes allows the diagnosis of definite RHD; whereas, if found in isolation, borderline (subclinical disease) RHD may be diagnosed. The guidelines are extensive, composed of many specific criteria, so we will abstain from reproducing them in their entirety here. Figure 2.6 exemplifies the usage of WHF's guidelines for the diagnosis of RHD in the echocardiogram of an 18-year old patient.

Prevention

RHD often remains asymptomatic until serious complications develop. However, prophylaxis in the form of regular penicillin injections can be initiated to prevent new ARF episodes and the disease's progression. During the later decades of the 20th century, the prevalence of RHD in high-income countries reduced significantly due to improvements in sanitation and medical follow-up [20]. However, areas where RHD remained endemic mostly present resource-poor settings with little to no medical records on ARF for the population.

The burden of RHD and its complications can be reduced by interventions during different stages of the progression from GAS infections to RHD:

Primordial Prevention. As ARF and RHD are predominantly diseases of social, environmental, and economic poverty, primordial prevention can be defined as improving socioeconomic and living conditions and having well-organized, effective health systems. This is arguably the most important and effective population-based strategy for the prevention of both ARF and RHD. Its effect could be seen during the 1940s and 1950s in countries such as Denmark [31] and the United States [48], even before the introduction of penicillin. Sustained political and economic change in LMICs is needed for this to happen in currently endemic areas.



(a) (b) (c)

Figure 2.6: **Diagnosis of Definite Rheumatic Heart Disease in the echocardiographic images of an 18-year-old obtained during a screening program. (a)** shows a > 2 cm jet of mitral regurgitation in parasternal long axis Doppler viewpoint; **(b)** and **(c)** show a > 2 cm jet of aortic insufficiency (yellow arrows) in parasternal long axis Doppler and apical 4 chamber Doppler viewpoints. Ao, aorta; LA, left atrium; LV, left ventricle.

Primary Prevention. Consists in the early diagnosis and treatment of GAS infections to prevent the first episode of ARF. Reliable and affordable methods for diagnosing and treating GAS infections are needed for primary prevention to be effective. Although definitive evidence for the efficacy of this intervention is still lacking, studies from the late 20th century in Cuba [79], Costa Rica [5], and the French Caribbean islands [7], suggest that the integration of this approach along with other economic developments contributed to the reduction of ARF in these countries. This is still a challenge in many LMICs, as there is a shortage of skilled staff and resources at the primary health care level. Also, many patients consider it a low priority to attend to a doctor with a sore throat [36].

Secondary Prevention. In patients who have had an episode of ARF or already have established RHD, long-term penicillin treatment, in the form of monthly injections, can prevent the development and worsen of RHD [121, 107]. Also, several studies have shown that even regression of the disease in 50-70% is possible over 10 years, especially for mild disease [97, 69]. Secondary prevention strategies seem feasible even in resource-limited settings, due to penicillin, in its many forms, being off-patent and generally an inexpensive antibiotic.

Tertiary Intervention. The aim of tertiary intervention of RHD is to reduce symptoms, disability, and premature death in patients who already have RHD-associated complications, therefore being a palliative treatment and not a prevention strategy itself. Treatments include medications to reduce heart failure and treat abnormal heart rhythms. However, the only definitive treatment is heart surgery to replace the damaged valves, which is not only costly but most of the time not available in endemic areas [57, 127]. In 2013 alone, the Brazilian Public Health System reported 5,169 hospitalizations related to ARF, and 8,841 related to chronic RHD, at a cost of 33 million USD, mostly related to cardiovascular surgeries [89].

2.1.3 Echocardiography

Echocardiography (echo) consists of sequential ultrasound images of a segment of the heart, taken with a probe that emits radiofrequency waves and receives them back. The signals are post-processed to form an image according to each wave's time to reflect from the surface it encountered. Figure 2.7 shows a segment of the heart being scanned with a probe and generating an echo image. The probe is small enough to fit between

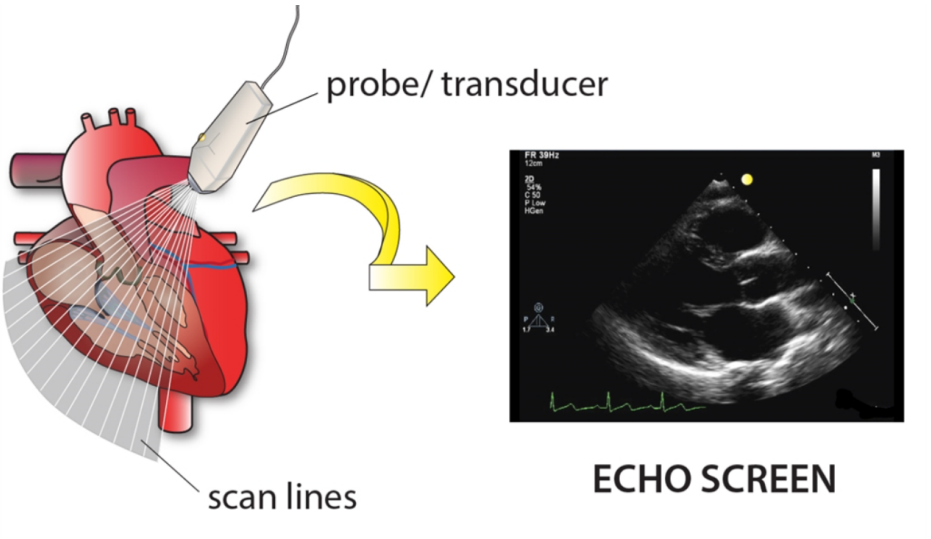
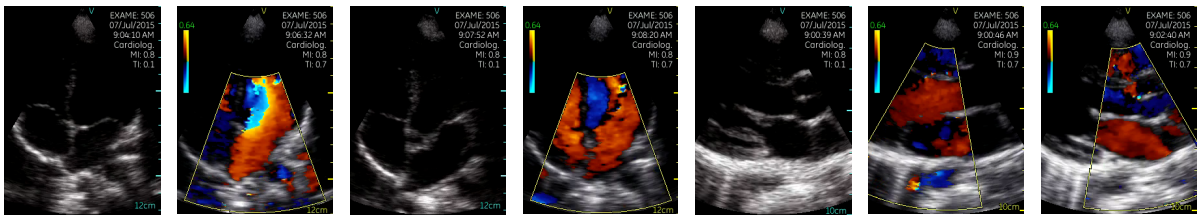


Figure 2.7: **Ultrasound probe scanning a segment of the heart to generate an echo image.** Image reproduced from [36].

the patient's ribs to prevent bone interference in the imaging process, and positioning the probe in different positions results in different cardiac viewpoints (views).

Echo is non-invasive, relatively inexpensive, and easy to use, making it the most widely adopted cardiovascular imaging modality [81]. Through multiple viewpoints and different modes, it is capable of providing rich information on the size, shape, and function of the heart, being crucial for diagnosing a range of heart conditions and reducing CVD-related deaths [33, 88, 81, 126].

Even though there are multiple echo modes, in this work we focus only in two: B-Mode and Color Doppler. Also, only 3 different viewpoints are used due to their relevance for RHD diagnosis: Apical 4 Chambers, Apical 5 Chambers, and Parasternal Long Axis. Figure 2.8 exemplifies the combination of modes and viewpoints used in this work.



(a) (b) (c) (d) (e) (f) (g)

Figure 2.8: **Examples of echocardiogram viewpoints and modes used in this work.** Frames were sampled from different videos of the same patient.**(a)** Apical 4 Chambers **(b)** Apical 4 Chambers with Doppler **(c)** Apical 5 Chambers **(d)** Apical 5 Chambers with Doppler **(e)** Parasternal Long Axis **(f)** Parasternal Long Axis with Doppler on the Mitral Valve Level **(g)** Parasternal Long Axis with Doppler on the Aortic Valve Level.

B-Mode

B-Mode (brightness mode), also called 2D echo, forms the basis of echocardiography. The amplitude of the wave determines the brightness of each pixel in the image returned to the probe, which allows real-time tissue visualization and, therefore, the analysis of anatomical structures, stationary or at motion. In RHD diagnosis, according to [88], this mode is used mostly for detection of morphological abnormalities in the mitral and aortic valves, e.g., chordal thickening or uncommon leaflet motions. B-Mode echocardiograms are exemplified in Figures 2.8a, 2.8c, and 2.8e.

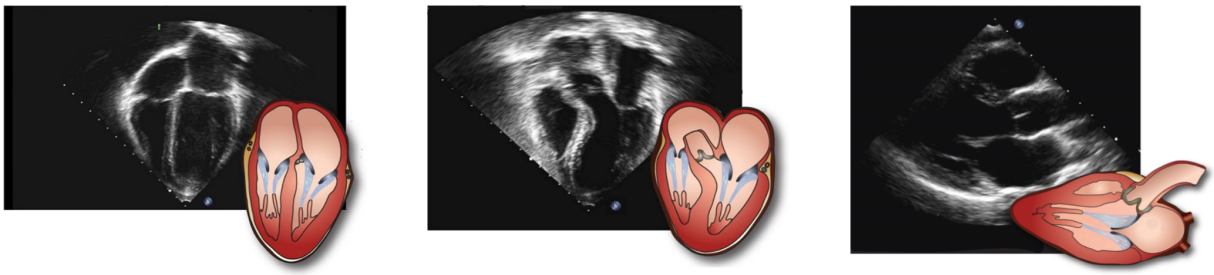
Color Doppler

Color Doppler mode can determine, with certain limits, the speed and direction of blood flow by utilizing the Doppler effect along with ultrasonography. Doppler information in a region of interest (ROI) is encoded in a color scale and then superimposed on B-Mode echo images. Blood flowing away from the probe is depicted in blue by convention (red on the opposite), with color intensity varying according to flow speed.

This echocardiography mode is very important for RHD diagnosis, as abnormal blood jets can be directly seen and then measured at their peak to assess if they are physiological or pathological [88]. Color Doppler echocardiogram are exemplified in Figures 2.8b, 2.8d, 2.8f, and 2.8g. The color scales can be seen in the top left of each image and their ROI is within the yellow borders.

Cardiac Viewpoints

Transthoracic echocardiography, which is the most conventional type of echo, consists of multiple videos of the heart's views, which are obtained by placing the probe in standardized places on the patient's chest. Having more than one view is important for RHD diagnosis, as the criteria in WHF's guidelines establish that for valve regurgitation to be pathological, it should be present in at least two views, having a size above a threshold in at least one of them [88]. Figure 2.9 shows the cross-sections of the heart that should be



(a) Apical 4 Chambers (b) Apical 5 Chambers (c) Parasternal Long Axis

Figure 2.9: **Cross-sections of the heart related to the cardiac views used in this dissertation.** Image adapted from [36].

probed in order to obtain the viewpoints used in this dissertation. It is worth noting that the Parasternal Long Axis (PLAX) viewpoint presents a clear view of the entire left side of the heart and the start of the aorta, making regurgitation detection and measurement easier.

2.1.4 Screenings

In epidemiology, screening is a strategy to assess the presence of preclinical disease in individuals or a specific population, aiming at intervening to improve health outcomes through further investigation and treatment of identified conditions. The WHO first published guidelines for screenings in 1968 [119], which were revised in 2008 with the emergence of new technologies [4].

As described in Section 2.1.2, secondary prevention strategies seem the most promising for regions where RHD is endemic, so screenings aim to identify the disease in its latent (subclinical) stage, intervening with prophylaxis to prevent its progression. As there is often time after an initial ARF episode and the development of advanced cardiac disease, which can occur as late as decades later [18], most screening programs have their demographics set to children and young adults.

The first study on echocardiographic screening for RHD was published in 1996 by [3] and demonstrated that the use of echocardiography detected more children with RHD than auscultation, which was the screening method previously used. However, at the time, the technological capabilities of portable echocardiograms were severely limited. Since 2004, the WHO has recommended screening in high-prevalence regions [19]. According to the 2020 Population Reference Bureau's estimates, approximately 84% of the children under 15 years of age (around 2 billion) live in countries endemic for RHD and are potentially at risk of developing the disease [85].

The advent of portable echocardiographic technology, around the size of a large laptop computer, made this a valuable diagnostic tool more widely available in resource-poor settings and remote locations, thus transforming the diagnosis of both ARF and RHD [32]. This was followed by the release of WHF's evidence-based standardized guidelines with echocardiographic criteria to facilitate the early diagnosis of RHD [88]. Consequently, over the past decade, echocardiographic screening for RHD has emerged as a potential strategy for the global control of the disease [77]. However, it is evident that the human, infrastructure, and financial resources required for testing, diagnosis, follow-up, monitoring, and quality assurance will be considerable and are likely to be a major challenge in resource-limited settings [100].

Figure 2.10: **Pocket-sized ultrasound device with human hand for scale.** This is a Vscan Extend™ handheld ultrasound device from GE Healthcare.

Even though portable echo machines are much cheaper and more portable than traditional hospital-based equipment, newer developments in ultraportable handheld ultrasound machines offer even greater portability at a much lower cost. Figure 2.10 depicts a pocket-sized Vscan Extend™ echo device with a human hand for scale, which was the device used during the screenings that generated the data used in this dissertation. These devices are capable of producing 2D and color images, are simple to use by inexperienced staff [88], and more recent versions are using smartphones, offering greater reach for screening teams at reduced cost [9]. However, they lack the spectral Doppler imaging capabilities needed to use the WHF criteria fully. Nonetheless, [11] showed that when used by experienced cardiologists with modified WHF criteria, an encouraging accuracy compared to standard portable echocardiography was obtained. The same research group reported that handheld echocardiography is more sensitive than auscultation [45].

Screening surveys or population-based screening programs require considerable organization and human resources. LMICs are evolving strategies such as task-shifting [118] to trained non-experts to make this vital strategy available and affordable for the most remote and poorest populations [35, 78]. Initial studies found that training non-experts

nurses or medical students to perform echocardiography for RHD screening was feasible [27, 8, 99]. Subsequent studies evaluated the possibility of task-shifting even the diagnosis through simplified protocols to trained non-experts [73, 84, 12, 35]. In screenings with non-experts, they were instructed to analyze the color Doppler echos while in the point-of-care facility to identify possible mitral and aortic regurgitations, immediately referring patients to cardiac centers according to specific criteria. Exams were later evaluated by experts. In one case, a telemedicine setup was used, with research cardiologists in Brazil and the United States reporting diagnosis through a remote system [76].

Efforts to reduce screening costs continue. Future directions for the evaluation, diagnosis and prevention of RHD worldwide are stated in Chapter 5 of [32]:

> The future for the clinical evaluation and diagnosis of RHD around the world lies in the development of affordable, accessible echocardiography with task-shifting of screening and diagnosis. These need to be employed effectively in hyperendemic communities to empower them to care for their people and improve knowledge, experience, and outcomes for future generations

We argue that using computer-aided diagnosis systems for automated identification of RHD during (referrals in the point-of-care facility) or after screenings can help make screenings even more widespread. The system could be made available through a cloud-based application for telemedicine or embedded directly into screening devices.

2.2 Computer-Aided Diagnosis

Computer-aided diagnosis (CAD) systems are responsible for assisting clinicians in the diagnostic process. CAD systems may use diagnostic rules to emulate the way a skilled human expert makes diagnostic decisions, but more advanced systems are able to analyze clinical data and learn from patterns to infer the diagnosis. Systems that can improve automatically by learning from data are said to perform Machine Learning (ML).

Since the late 1950s, due to advancements in computing technology, biomedical researchers began exploring the possibility of using computers to investigate and solve problems in biology and medicine. Some of those studies were ultimately directed to the development of systems for computer-based medical diagnosis [62, 110, 117]. However, by the early 1970s, it became clear that there were some severe limitations in delivering accurate diagnosis when using traditional methods such as flow-charts [15], statistical pattern-matching [92], or probability theory [49]. Early expectations of the potential of the newly emerged computer-based approaches tended to be overly optimistic. At the

beginning of these early studies, researchers were hoping to develop entirely automatic computer-aided diagnostic systems, but new, more sophisticated methods were necessary.

Even though CAD systems, in the form of artificial neural networks, have been used to analyze 2D echocardiograms since 1990 [24], the past decade has witnessed an explosion of successful classification approaches in many areas of medicine thanks to the development of Deep Learning (DL), a subfield of ML. Historically, constructing a ML system required domain expertise and human engineering to design feature extractors that transformed raw data into suitable representations from which a learning algorithm could detect patterns. In contrast, DL is capable of not only learning the patterns themselves, but also the best way to extract features from the data in which patterns appear more clearly.

The components that comprise DL are not new, but recent increases in computational power and the availability of massive datasets allowed performing more complex learning in treatable time. Artificial neurons form the basis of DL.

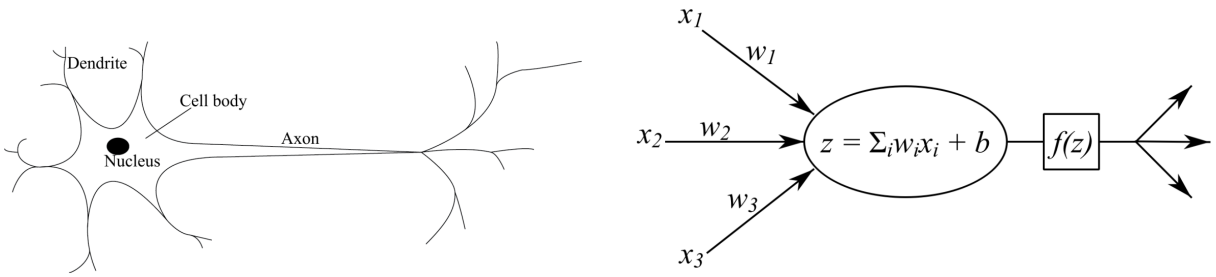
2.2.1 Artificial Neurons and Neural Networks

Artificial neurons are the basic building blocks of neural networks and, as the name suggests, are inspired by the biological neuron. In general, biological neurons consist of dendrites, a cell body with a nucleus, and an axon. The cell body receives electrical signals through the dendrites, the signals are processed by the cell body, and a response is transmitted through the axon. Similarly, an artificial neuron receives an input signal, processes it by summing the weighted inputs, and outputs the result. The output is given by the following equation:

$$z = \sum_i w_i x_i + b$$

where z is the output, w is a weight, x is input data and b is a correction bias. This mathematical description for an artificial neuron is also called a perceptron since it was proposed by [93]. Similarities between a biological and an artificial neuron are drawn in Figure 2.11.

We can stack multiple neurons to receive multiple inputs at once. If we connect all of the outputs of our neurons to the inputs of new ones added alongside, we will create a fully-connected feedforward Artificial Neural Network (ANN) with two layers, commonly called a Multi-layer Perceptron (MLP). The first layer of a neural network is the input layer, and the last layer is the output layer. The layers in-between are called hidden



(a) Biological neuron (b) Artificial neuron

Figure 2.11: **Biological and artificial neurons.**

layers. The total number of layers in an ANN is referred to as the depth of the network, and the process of training multilayered ANNs is referred to as deep learning.

A simple perceptron can only learn linear mappings of the input data [46]. To be able to learn more complex non-linear mapping, we can process the information that is propagated between layers of an ANN with a non-linear activation function. The most common activation function used is the rectified linear unit function (ReLU) due to its proven performance improvement [46]. That way, a MLP with two layers is already capable of solving the XOR in 2 dimensions, which is not linearly separable [96]. Figure 2.12 shows how, as data flows through the layers of an ANN, the input space becomes warped, and data points become increasingly distinguishable. Following this principle, highly complex functions that map any input to arbitrary labels can be learned. It is possible to approximate any computable function with feedforward ANNs, as shown by [52].

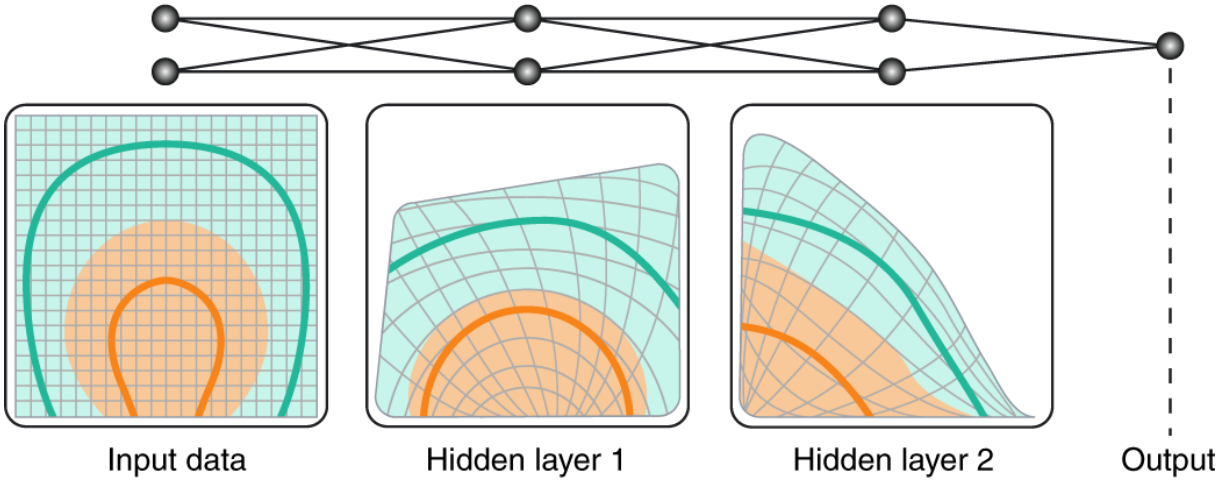


Figure 2.12: **Neural network layers making data linearly separable.** Different input classes are denoted by different colors. Image adapted from [39].

2.2.2 Deep Neural Networks

A Deep Neural Network (DNN) is an ANN with multiple hidden layers. Typical DNNs contain millions of trainable parameters (weights), which usually require large amounts of data to be adjusted (trained) properly. Figure 2.13 exemplifies a large-scale network receiving multiple types of data at once. Hidden layers 1 and 2 work as feature extractors, while hidden layers 3 and 4 combine all the features and classify the input into one of three classes.

DNNs are trained to minimize the error, or the loss, between the label attributed to a sample and the network's output after receiving that sample as input. The error value is calculated according to a loss function, and the network parameters are subsequently adjusted based on how much each sample contributed to the loss. First, the gradient of the loss function concerning each parameter is calculated using the backpropagation algorithm, initially described in [95]. With the gradients at hand, adjusting the weights to minimize the loss consists of solving a non-convex optimization problem, meaning that the network may go through local optima during the process. The most commonly

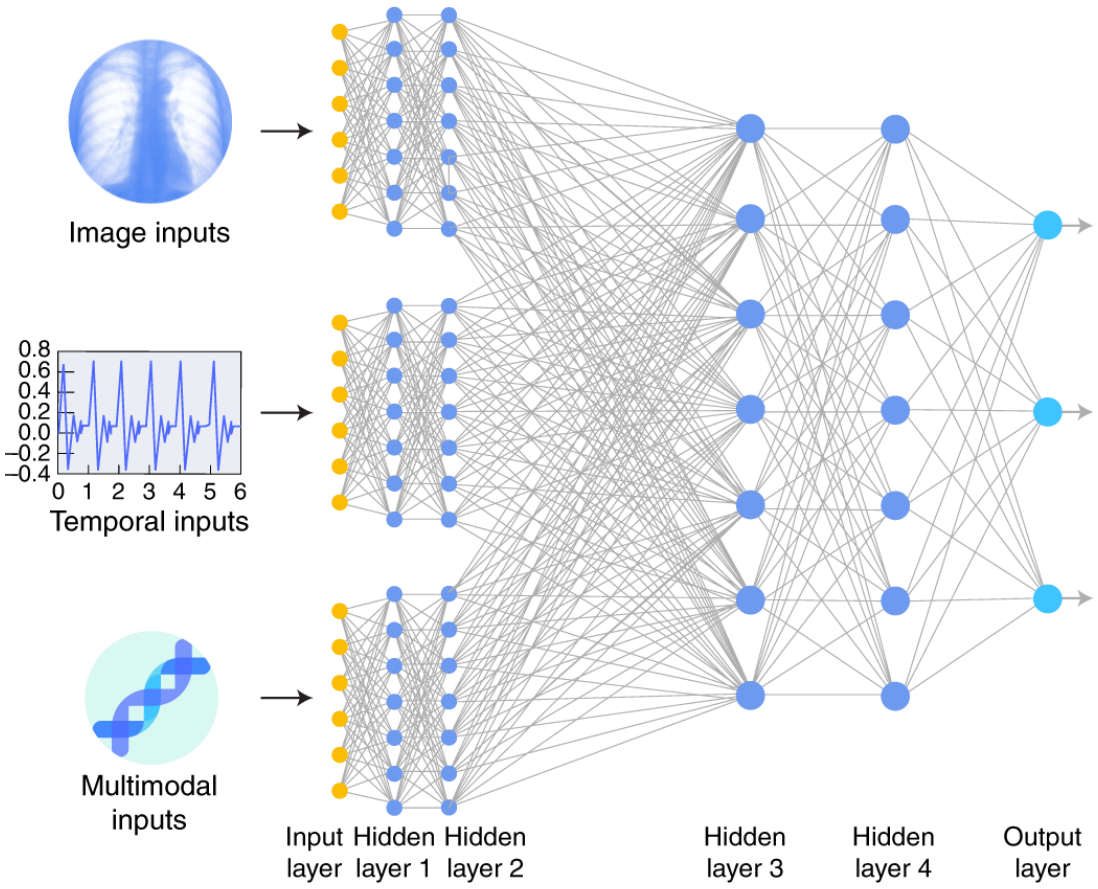


Figure 2.13: **Example large-scale network that accepts a variety of data types as input.** Image adapted from [39].

used optimization algorithm is the Stochastic Gradient Descent (SGD) [46]. This is an iterative process, where typically training samples are seen multiple times across epochs. An epoch in ML consists in passing through the entire training dataset once. With these mechanisms, DNN systems can featurize and learn from a variety of data types, with different network architectures being more tailored for specific types of data.

2.2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are DNNs with an architecture specialized for processing data with a grid-like structure[46] and were first inspired by visual cortex research done by [54]. CNNs can efficiently work with one-dimensional (time-series) data, two-dimensional images, or three-dimensional volumes such as videos or point clouds. They are a powerful tool for image classification and regression problems, having grown to be central in the field of computer-aided diagnosis [39].

In the context of medical imaging, CNNs receive as input a matrix of pixels representing an image or a volume of pixels representing a video, which are sequentially downsampled by convolutional layers. Convolutional layers consist mainly of three operations: convolution, non-linear activation, and pooling. The convolution operation filters the input data, generating feature maps that preserve multi-dimensional information since the activation of each learnable filter (kernel) is dependent on neighboring inputs. This information is spatial when using 2D CNNs and spatio-temporal for 3D CNNs. Non-linear activation, e.g., ReLU, is applied to the feature maps to warp the representation into iteratively more distinguishable features. Finally, the pooling of the feature map is performed to reduce the feature representation, effectively downsampling at each step. Kernels have weights that are adjusted according to the minimization of a loss function, as described in the previous section.

Figure 2.14 is a typical configuration of a CNN architecture for medical imaging diagnosis. The configuration mainly consists of an input image, followed by a sequence of convolution operations joint with a non-linear activation function and pooling function responsible for feature extraction. The extracted features are then fed to fully-connected feedforward layers, which act as a classifier. The number of neurons in the output layer equals the number of different labels in the classification task. A softmax function is used so that the outputs represent the probability of the current input representing each class, sometimes referred to as the model's confidence.

The first real-world application of CNNs with significant accuracy was the recognition of hand-written digits [61]. However, the breakthrough for this technology happened

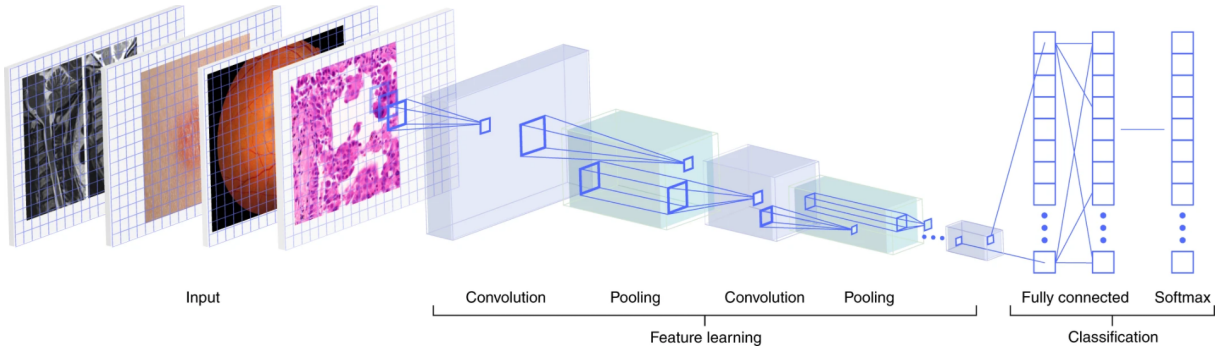


Figure 2.14: **Medical imaging automated diagnosis using a typical architecture with Convolutional Neural Networks.** Image reproduced from [39].

only in 2012, when [59] improved the performance of image classification on the ImageNet Large-Scale Visual Recognition Challenge by almost 100% using a CNN. One of their main contributions was a very efficient method for training DNNs using graphic processing units (GPUs), which sparked remarkable interest in DL among the ML community, subsequently leading to significant advancements in the field.

Since them, CNNs have been developed to detect diseases in many medical fields [98], such as pneumology [53, 86], optometry [43, 124, 87], and oncology [112, 38]. They have even achieved physician-level accuracy at a broad variety of diagnostic tasks, including detecting anomalies in optical coherence tomography [29] and identifying moles from melanomas [38]. Clinics are beginning to employ object detection and segmentation in images for urgent and easily missed cases [123]. The primary limitation when building a supervised CAD system for a new medical imaging task is access to a sufficiently large, labeled dataset [123].

2.3 Related Work

Related DL literature mainly comprises studies on echocardiogram viewpoint identification [42, 67, 68, 125], heart chamber segmentation [125, 21], and classification of heart disease [65, 68, 44, 125]. Viewpoint identification is primarily made with regular frame-based 2D CNNs, with architectures similar to [59]. Works on heart chamber segmentation mostly use models equivalent or derived from U-Net [91]. Classification of heart diseases through DL methods has been, in most cases, used for the detection of chamber hypertrophy or dilation. Automatic identification of RHD through conventional echocardiogram exams was not previously addressed in the literature to the best of our knowledge.

In [65], left ventricle diameter, left atrium area, and interventricular septum width

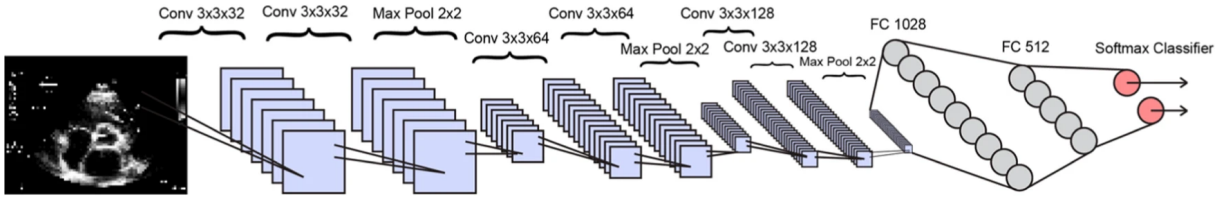


Figure 2.15: **Example of regular frame-based convolutional neural network used for cardiac disease diagnosis.** Image adapted from [67].

are estimated from selected frames of end-diastole through a regression network that receives 3 different viewpoints (Apical 2 Chambers, Apical 4 Chambers, and Parasternal Long Axis) in individual convolutional modules. Through the concatenation of the learned feature maps, followed by more convolutional layers, a fully connected module outputs the 3 measures for each viewpoint, which, if above certain thresholds, can be indicative of hypertrophic cardiomyopathy (HCM) or dilated cardiomyopathy.

The work by [68] compares results for binary classification of left ventricular hypertrophy (LVH) between two proposed architectures. The first architecture consists in two image segmentation phases performed by U-Nets, interleaved with phases of viewpoint and LVH classification performed by either ResNet50 [51] or VGG-16 [101] networks that have been pre-trained on the ImageNet dataset [30]. The second proposed method is an ensemble of 3 GAN models trained in a semi-supervised fashion, which shows a superior accuracy in the detection of the condition. The semi-supervised architecture was proposed to take advantage of an unlabeled set of images with 33 times the number of labeled ones. Both architectures received only 2 frames from each Apical 4 Chamber video as input to ensure that only the diastolic phase is considered.

[44] performs a binary classification of LVH and enlargement of the left atrium. Their proposed 75-layer CNN, named EchoNet, was based on the architecture of Inception-ResNet-v1 [105] and received 20 frames from each echocardiographic video that depicted an Apical 4 Chambers viewpoint. Final predictions were obtained by averaging all the predictions from individual frames.

Finally, [125] uses a VGG-13 [101] to diagnose HCM, cardiac amyloidosis (CA), and pulmonary artery hypertension (PAH). Separate networks were derived for Apical 4 Chambers and Parasternal Long Axis images for HCM and CA, and only a single Apical 4 Chambers network for PAH. Each model was trained with 3 random images from each video, and the accuracy of the model was assessed by an average of the probabilities output for 10 random images of an echocardiogram, with further aggregations of the results by a mean of videos per study, and a mean of the trained models in the case of HCM and CA.

All of the described related works use regular frame-based CNN architectures that disregard temporal information, as exemplified in Figure 2.15. When considering the morphological and functional criteria for RHD diagnosis established in [88], some structural features, such as Chordal thickening of the mitral valve or Irregular/focal thickening

of the aortic valve, can be identified in single frames if they are not obstructed by other structures or blood flow. However, the identification of most morphological and functional features used for diagnosis, e.g., Excessive mitral valve leaflet tip motion during systole or Pathological aortic regurgitation with jet length ≥ 1 cm, is directly correlated to the spatio-temporal observation of some heart structures, and, therefore, a temporal-aware method seems more suitable for this specific task. Also, it is worth noting that none of the related works use data acquired from echocardiographic screenings. In all of them, experts performed exams with gold standard equipment and in controlled hospital environments, thus probably providing data with superior quality to the machine learning models.

2.4 Summary

This chapter described the background and related works relevant for this dissertation. We explain the important clinical concepts regarding RHD's symptoms, causes, diagnosis and prevention, also highlighting the importance of increasing the availability of screenings to reduce the burden of RHD in the developing world, something that can be benefited by the development of applications for automated diagnosis of the disease. We also briefly describe advancements in computer-aided diagnosis for many areas in medicine, describing how the main method behind this revolution, CNNs, works. The chapter finishes by listing related work in computer vision applied to echocardiography. There are no previous works in the identification of RHD in conventional echocardiograms, and the related work in disease classification always use single frame classification, something that we hypothesize to be suboptimal for the task we are tackling.

The next chapter describes our first proposed method for the automatic diagnosis of RHD, which not only encodes temporal information present in echocardiograms but also generates final predictions in a more sophisticated way. The method is compared to the performance of a frame-based approach similar to the ones used in [125] and [68].

Chapter 3

Supervised Diagnosis with Temporal-Aware Learning

The task of identifying structural cardiac abnormalities has been shown to be efficiently done through single frame classification, as described in the last chapter. However, the diagnosis of Rheumatic Heart Disease (RHD) also relies on observations that span multiple sequential frames of an echocardiogram. We, therefore, hypothesize that a temporal-aware method would perform better in such a task.

In this chapter, we present a method that uses a 3D convolutional neural network (CNN), C3D [108], for individual video classification, therefore encoding temporal information. The method is also composed of a supervised meta-classifier based on Random Forest [17] to aggregate predictions from the previous classifier regarding the same exam into a final diagnosis for RHD. We start by describing our dataset, as its nature directly impacts architecture decisions. This chapter is based on [72].

3.1 Dataset

Our dataset comprises 11,646 echocardiographic videos in MP4 format (resolution 320 Œ 240 pixels), taken with Vscan Extend™ devices (GE Healthcare, Milwaukee, WI, USA), which sum up to 912 complete exams of unique patients. The data was acquired as part of screening programs in Uganda [11, 84] (359 exams) and the PROVAR screening program in Brazil [76] (553 exams). The programs were conducted between 2012 and 2016 and screened children attending public schools. It focused on the early detection and prevention of disease progression, and screenings were performed mostly by trained non-experts (584 exams). Table 3.1 presents the demographic profile of our dataset. Note that only a subset of 528 exams, all from the PROVAR study, have complete demographic data annotated. The observed discrepancy in the prevalence of rheumatic valve disease by gender, with a remarkably higher prevalence in females, is also noted in other stud-

Table 3.1: **Demographic data of subjects present in the dataset.**

	Overall ($N = 912$)	**RHD Negative** ($n = 456$, 50%)	**Borderline RHD** ($n = 349$, 38.3%)	**Definite RHD** ($n = 107$, 11.7%)
Patient demographics, *n* (%)	528 (100%)	265 (50.2%)	231 (43.8%)	32 (6%)
Gender, *n* female (%)	316 (59.9%)	145 (54.7%)	150 (64.9%)	21 (65.6%)
Age, years (SD)	13.1 (3.1)	12.6 (3.1)	13.6 (3.0)	13.1 (3.4)

ies [76, 104, 109]. The studies were approved by the institutional review boards of both the Childrens National Health System and Universidade Federal de Minas Gerais. In both studies, informed consents were collected during visits to schools. After the intervention, a letter explaining the study procedures was sent to families with the consent and assent terms. Patients were only included after returning the signed consents, and their echocardiograms were de-identified.

The estimated RHD prevalence in the examined regions and age group is $\leq 4.2\%$ [76, 11]. However, due to the sensibility of the evaluated learning methods to extreme imbalances in the distribution of labels [114], the dataset comprises 456 (50%) RHD negative and 456 (50%) RHD positive exams, which are composed of Borderline RHD and Definite RHD diagnosis.

Each exam contains, on average, 12.77 (3.59) videos, each possibly representing one of seven different viewpoints of the patient's heart. The viewpoints are depicted in Figure 2.8 include Apical 4 Chambers with and without Doppler, Apical 5 Chambers with and without Doppler, Parasternal Long Axis, Parasternal Long Axis with Doppler on the Mitral Valve Level, and Parasternal Long Axis with Doppler on the Aortic Valve Level. Videos in the dataset are missing the labels for the viewpoints they represent. Figure 3.1 shows the distribution of number of videos per exam by video type (with Doppler or without Doppler). The vast majority of our videos contain Doppler, accounting for 69.77% of all the data, even though they represent only 57.11% ($\frac{4}{7}$) of the viewpoint classes.

For the PROVAR exams, five cardiologists with expertise in RHD examined all morphological and functional changes in mitral and aortic valves according to the WHF criteria. Two readers independently reviewed all abnormal echocardiograms, and discrepancies were resolved by consensus between three readers. The inter-reviewer agreement was 0.89 (95% CI 0.860.92), and the between-reviewer agreement 92% [76]. A similar reviewing process was executed for exams performed in Uganda. The self-agreement ranged between 71.4 and 94.1% (κ 0.470.84), and the between-reviewer agreement ranged from 66.7 to 82.8% (κ 0.340.46) [11, 84].

When considering the usefulness of the collected data for computer-aided diagnosis, the measures taken to make screenings more widespread pose some challenges. Handheld devices present a poor signal-to-noise ratio, which is exacerbated even more as the environments where screenings take place are often improvised, with substandard

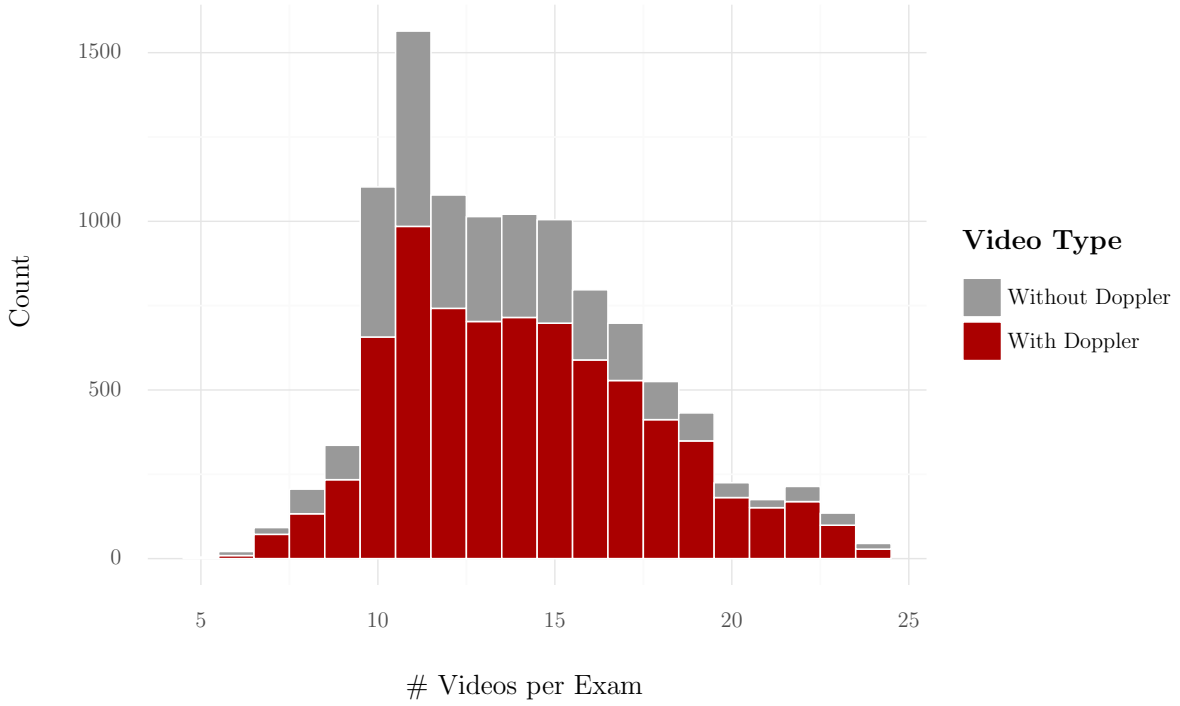


Figure 3.1: **Distribution of the number videos per exam by video type.**

infrastructure. Pediatric imaging has the potential to aggravate these errors even further due to the smaller size of the hearts, higher heart rates, and a limited ability to have the patients voluntarily hold their breath.

Videos can begin at any point in the cardiac cycle and may not even present a complete cycle in some cases. They have less than 2 seconds in most cases, being recorded at 12 frames per second and averaging 17 frames each. Another point that should also be taken into account is that the average number of videos per exam is well above the count of unique viewpoint classes, and this happens due to two behaviors reproduced by the professionals that performed the exams. Videos with recording problems, e.g., bad probe positioning or suboptimal imaging setup (skewed image gain), were not deleted, and, without any tag to differentiate and remove them afterward, noisy instances populate our final dataset. Besides, some videos did not correspond directly to any of the specified viewpoints. This happens because when a point of interest that would later help identify the presence of RHD is perceived, another video, zooming into the area, is recorded, changing the scale of cardiac structures to an unknown pattern.

3.2 Methodology

This section introduces the two main components of our proposed methodology: i) a deep CNN to classify the videos as RHD positive or negative and ii) an aggregation strategy, which accounts for the results of all videos of a single patient, as shown in Figures 3.2 and 3.3, respectively. The methodology starts by feeding a 3D CNN, i.e., C3D [108], with videos from all viewpoints of the patient exam. Then, the networks' outputs for all videos of a single patient are combined using a meta-classifier, as detailed next. It is noteworthy that information regarding the patient itself is not provided during the training phase, i.e., the CNN does not know which videos correspond to the same exam.

3.2.1 Convolutional 3D Network

We use the C3D as the backbone network of our method, as illustrated in Figure 3.2. The C3D network is a deep CNN that can learn from the temporal information by applying three-dimensional convolution operations. The network receives a tensor of 112 Œ 112 pixels Œ 3 color channels Œ 16 frames. The initial 16 frames of each video are used to train the network. Since some videos contain less than 16 frames, we add padding frames that are a balanced number of duplicates of the first and last frames until the required length is achieved. In a transfer learning fashion, we used the model pre-trained on the Sports-1M dataset [58]. Initially, visual features are extracted by convolution layers with small 3 Œ 3 Œ 3 kernels combined with the max-pooling operation. These features

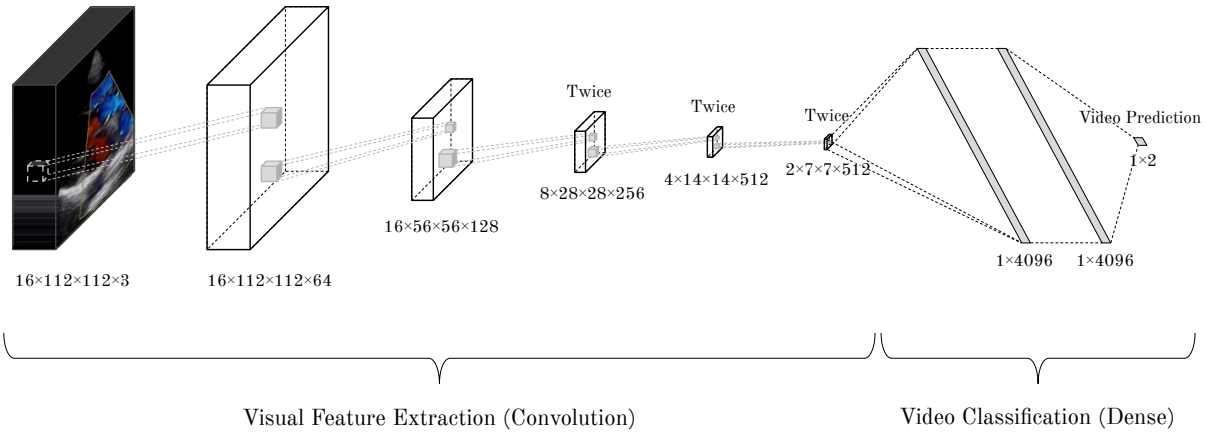


Figure 3.2: **C3D network architecture for video classification.**

are then fed to a fully connected set of layers, with the last layer composed of only two neurons and the softmax function as activation, outputting the probability of the video belonging to one out of two classes: RHD Negative or RHD Positive. In order to simplify the problem, the Borderline RHD and Definite RHD diagnosis were grouped into a single class, named RHD Positive. All other layers use ReLU as the activation function. To prevent overfitting and improve generalization, dropout [102] with a probability of 0.5 is implemented within the first two dense layers.

The C3D model minimizes the binary cross-entropy loss function $\mathcal{L}$ as follows:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=0}^{N} \left(y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \log(\hat{y}_i)\right),$$

where y_i is the label of the i-th sample, $\hat{y}_i$ the predicted probability of the positive class, and N the number of samples. In summary, the network tries to minimize the distance between the confidence in its predictions and the true diagnosis for each echocardiogram.

3.2.2 Aggregation with a Supervised Meta-Classifier

In the first step of our methodology, images from an exam are given to C3D independently. Next, the output of the CNN can be used in different ways to provide a diagnosis to a single patient. A standard approach to aggregate the results of all frames is to use a majority vote strategy, where each predicted video class counts as a single vote. However, a binary view of each prediction (positive RHD or negative RHD) disregards a great deal of information that could be useful for counterbalancing biases that emerged during the training of our model, therefore possibly improving the accuracy of our prediction. Thus, we propose a more sophisticated aggregation strategy that uses a supervised classifier to predict the diagnosis (see Figure 3.3). This aggregation strategy is based on a set of meta-features extracted from the probability distribution output by CNN over videos per exam, namely, the mean, standard deviation, skewness, and kurtosis.

Stacked generalization [120], now commonly referred to as stacking, is one of the most used ensemble learning techniques in machine learning. It combines multiple classification or regression models via a meta-classifier or a meta-regressor that leverages the output of the base models to give a final prediction. In the context of our study, there is only one base predictor but multiple instances that should be aggregated into a single output. As the number of videos per exam is not fixed, we use the statistical moments of our classifiers confidence distribution as inputs for the meta-model. The proposed aggregation strategy is agnostic to both the base classifier and the meta-classifier, as long as

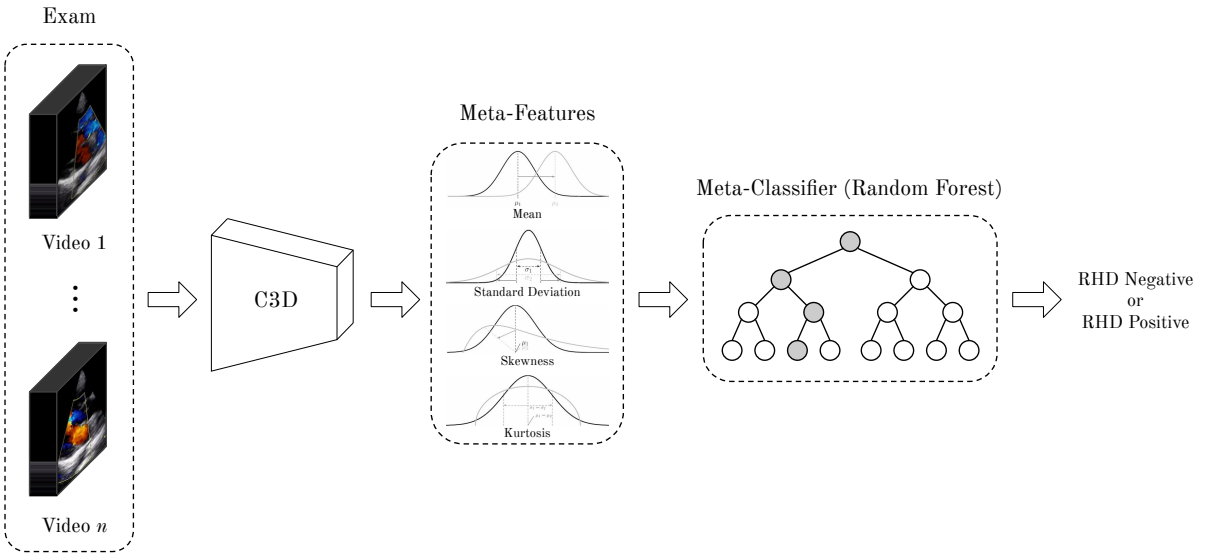


Figure 3.3: **Proposed supervised meta-classifier for result aggregation toward exam classification.**

the first can output its prediction as a probability. The meta-classifier of choice for this experiment was the decision tree-based Random Forest [17], due to its notorious efficacy when little is known about the domain being evaluated [40].

3.3 Experiments

3.3.1 Baseline, Hyperparameters and Implementation

We consider that one of the novelties of the methodology proposed here is the use of a 3D CNN to process videos as inputs instead of a 2D CNN, which works with images. Hence, to start with, we compare the proposed methodology against a frame-based method that uses VGG-16 as the backbone neural network. The VGG-16 [101] architecture is similar to the ones used in two related works [125, 68] and it is well-established in the computer vision community. VGG-16 is a 2D CNN that receives as input a still frame of 224×224 pixels $\times 3$ color channels. Following a similar methodology as the one in [125] we have sampled 10 random frames from each video to create instances for the network, and the network predictions are then aggregated using a majority vote strategy per video and then per exam, giving preference to the positive class in case of a draw. The model used was pre-trained on the ImageNet dataset. The loss function used

was also the binary cross-entropy.

Next, to measure the contribution of the proposed meta-classifier, we have also run experiments where the C3D results were aggregated using a majority voting to give the patient's final diagnosis. We have trained in our dataset the VGG-16 network (pre-trained in the ImageNet dataset) with the Adam optimizer, a learning rate of $1e-5$, batch size of 32 and 25 epochs, using early stopping with a patience value of 10. For C3D we used an SGD optimizer, a learning rate of $1e-3$, batch size of 16 and 25 epochs also, but with 5 as the patience for early stopping. The Random Forest model was trained with 200 estimators in the forest and a max depth of 75. Unlisted hyperparameters for all models were left to their default values. The set of hyperparameters for each method were chosen through a Random Search setup with 30 iterations for the neural networks and 500 for the Random Forest.

Our code was written in Python 3.6, and executed in a machine with Intel(R) Core(TM) i7-9700K CPU and an NVIDIA GeForce RTX 2080 Ti GPU. All the neural networks were implemented using Keras [23] with TensorFlow 1.12 [1] as the backend. The used Random Forest classifier is packed within version 0.20 of scikit-learn [83].

3.3.2 Experimental Setup

We have performed a binary classification with the Borderline RHD and Definite RHD diagnosis grouped into a single class, named RHD Positive. All echocardiograms were de-identified by applying a mask of black pixels to the area outside of the ultrasound beam during preprocessing, therefore omitting the metadata present in the images. As all echocardiograms in the dataset were collected using the same equipment and software, the size of this area was fixed. For the C3D inputs, videos were first rotated 90 degrees and then resized to $128 \times 171 \times 3 \times 16$. This was done to obtain a better aspect ratio when removing the mean cube of the original training data, a preprocessing step called whitening [14]. A centered cropping was then applied to generate the final data. As for the VGG-16, videos were directly resized to the expected input dimensions.

In order to train the model, tune hyperparameters and then diagnose new exams, we randomly split the dataset into training, validation, and test in an approximate 80:10:10 ratio. The splits were stratified, and videos from the same exam were always in the same data partition. Each patient has only one exam in the dataset. Hyperparameter tuning for both neural networks and the Random Forest meta-classifier used for aggregation was done using only the train and validation sets to prevent information leakage from the test partitions.

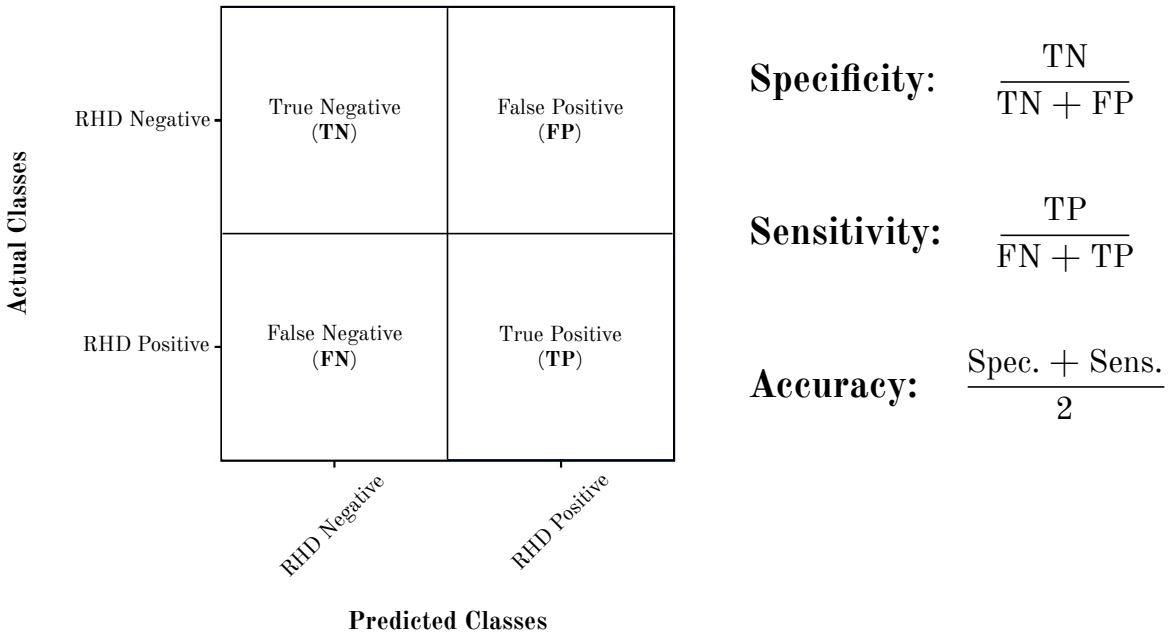


Figure 3.4: **Metrics used to assess models' performance and how to calculate them from a confusion matrix.**

The metrics used to assess models' performance were specificity, sensitivity and accuracy, calculated from a confusion matrix as described in Figure 3.4. The results displayed are obtained through a 10-fold cross-validation procedure, making each video go through the validation and test partitions only once. Folds are the same for all evaluated methods.

3.3.3 Results

Table 3.2 reports the mean specificity, sensitivity, accuracy for the test partitions in the 10-fold cross-validation procedure. We performed a Wilcoxon signed-rank test with 95% of confidence to compare the results of the three different methods, and the best method for each metric is highlighted in bold in the table. In cases where there is no evidence of statistical difference, both results are highlighted. The specificity and sensitivity obtained by the best model were 70.59% ($\pm$ 4.06) and 74.94% ($\pm$ 4.84), respectively. Averaged over the test partitions for each of the 10-folds, its accuracy was 72.77% ($\pm$ 3.49).

As expected, C3D with the majority vote is significantly better than VGG-16 for all metrics but specificity (where the results of both methods present no statistically significant difference) at the video level, showing the already stated importance of spatio-temporal information to the task at hand. Considering the exam level, which provides

Table 3.2: **Mean specificity, sensitivity, accuracy (with 95% confidence intervals) for RHD classification on the test set over a 10-fold cross-validation procedure for different levels of result aggregation.** Results in bold are the best for that metric according to a 95% confidence Wilcoxon signed-rank test. In cases where there was no evidence of difference, both results are highlighted. MV and MC stand for the Majority Vote and Meta-Classifier aggregation strategies, respectively.

Aggregation	Metric	VGG-16	C3D + MV	C3D + MC
Frame	Specificity	54.37 ± 2.59	—	—
	Sensitivity	56.90 ± 2.92		
	Accuracy	55.70 ± 1.12		
Video	Specificity	**59.59** ± 3.30	52.67 ± 5.67	—
	Sensitivity	55.17 ± 3.98	**67.71** ± 5.11	
	Accuracy	57.29 ± 1.46	**60.42** ± 1.66	
Exam	Specificity	**67.98** ± 5.30	57.92 ± 9.93	**70.59** ± 4.06
	Sensitivity	57.52 ± 5.12	**78.01** ± 6.48	**74.94** ± 4.84
	Accuracy	62.80 ± 1.11	67.95 ± 3.03	**72.77** ± 3.49

the final diagnosis, the proposed methodology significantly outperforms the other two methods with regard to accuracy, which is our final classification objective. The meta-classifier significantly outperforms the majority voting in terms of specificity, but there is no statistically significant difference for sensitivity. The considerable variance seen across methods for specificity and sensitivity can be explained due to the small number of videos and exams seen in the test partitions. At most, there are 46 exams and 621 videos for each class, which can be considered a small amount, given that our data already has high variability due to its nature of noisy acquisition, multiple viewpoints (with and without color Doppler), and different heart sizes.

An analysis to assess if Definite RHD cases are easier to identify than Borderline cases which is expected was also performed, and the reported sensitivities corroborate the expected results. In Table 3.3 we break the results for the RHD Positive class considering its original subclasses: Definite RHD and Borderline RHD. The sensitivity obtained for the Definite RHD class is comparable to the 83% (95% CI, 76%-89%) overall sensitivity achieved by non-expert users in RHD identification after following a computer-based 3-week training curriculum, as reported by [12]. More detailed comparisons with non-expert human performance are drawn in the next chapter.

Table 3.3: **Sensitivity values (with 95% confidence intervals) for exam classification of the two subclasses aggregated as RHD Positive in our dataset.**

Diagnosis	Subclass Exam Sensitivity
Borderline RHD	71.90 ± 5.19
Definite RHD	85.78 ± 6.41

Table 3.4: **Average meta-feature importance percentage observed across folds using the C3D network as the base classifier.**

Meta-Feature	Importance (%)
Confidence Mean	76.4
Confidence Std	6.3
Confidence Skewness	12.6
Confidence Kurtosis	4.7

3.3.4 Importance of the Meta-Classifier

Regarding the proposed aggregation strategy, it achieved significantly better results than the baselines. We took advantage of the decision tree method's interpretability to explore the meta-classifier's functionality further and compare its effects against the solo C3D model. Table 3.4 presents the average feature importance detected by the meta-classifier across folds, indicating that the distribution moments used as features were indeed relevant for a better prediction. An analysis of feature importance showed that all distribution moments were indeed relevant for a better prediction. For a simple comparison, training the same model only with the Confidence Mean feature, responsible for most of the feature importance, the cross-validation accuracy of the C3D + Meta-Classifier would be 70.47%, which is not statistically better than the C3D with the majority vote strategy (confidence of 95%).

Figure 3.5 shows the distributions of the four statistical moments for the training and test partition of the first fold of our data, discriminated by the correct diagnosis. The accuracy of the C3D + Meta-Classifier method in the test partition of this fold was 70%, with a specificity of 60% and a sensitivity of 80%. The Confidence Mean is the most prominent feature for separating the two classes, followed by the Skewness.

We assumed that the meta-classifier counterbalances biases acquired during the training of the base model. If this holds, results from a majority vote ensembling strategy should be more unbalanced in nature. Figure 3.6 shows the confusion matrices obtained for both aggregation strategies along with the C3D network. By comparing Figure 3.6b with Figure 3.6c one can observe a loss of sensitivity around 3 absolute percentage points with a compensatory increase in specificity of almost 13 absolute percentage points. This same pattern appears in executions with different hyperparameters. This indicates that the proposed aggregation strategy possibly identifies noisy videos in an exam through unexpected disruptions in the confidence distribution and filters the noise out, obtaining more accurate predictions overall.

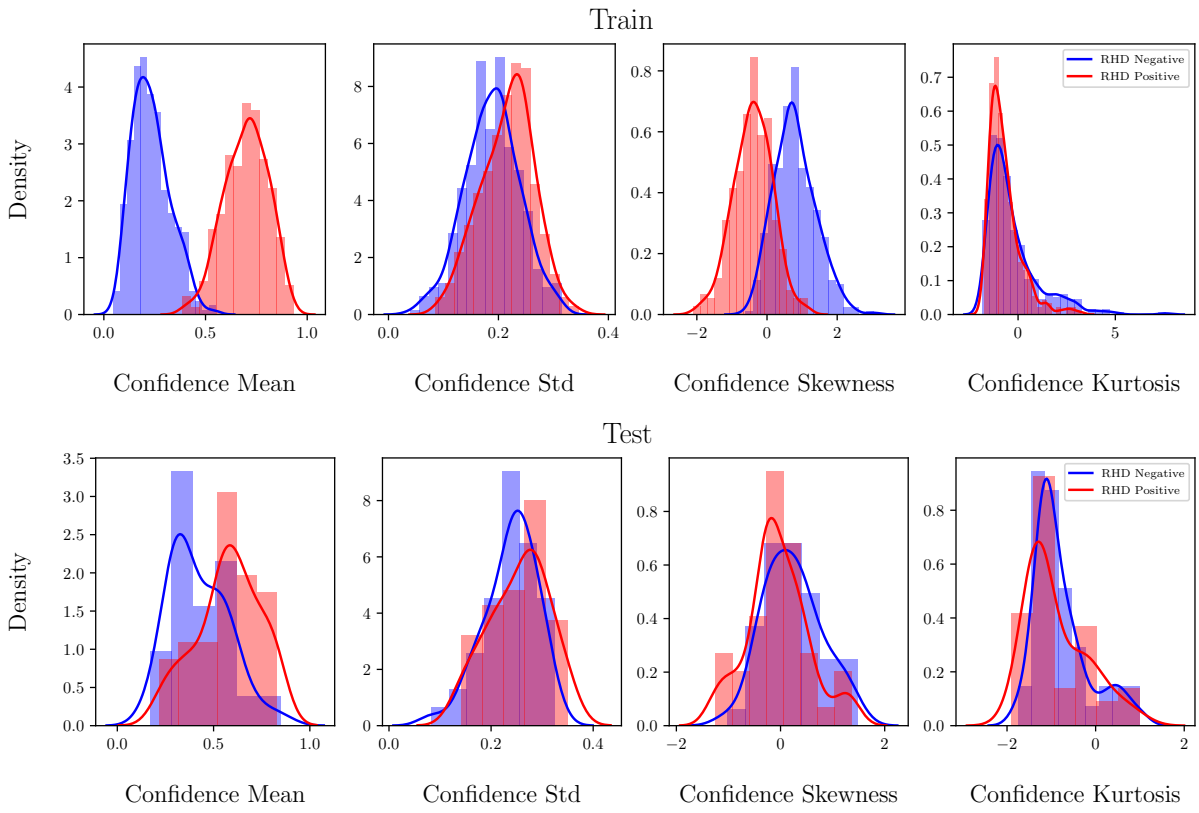


Figure 3.5: **Distributions of meta-features for the training and test partition of the first fold of our data considering a 10-fold cross-validation, discriminated by the correct diagnosis.**

3.3.5 Classification Examples

Figure 3.7 shows examples of four frames extracted from four videos where the proposed model classified an exam as RHD positive or negative with high confidence. The videos are from four different exams, and we have provided in Appendix A the detailed information used by our method during its decision making process on these exams when in the test set.

Analyzing the images classified as RHD positive or negative in Figure 3.7, we notice that images (a) and (b) have quality problems. In Figure 3.7a the blood flow from the abdomen (in blue) was captured by the Doppler, which probably confused the network due to a pattern similar to a valve regurgitation, and led to the classification of a negative example as positive. Figure 3.7b shows a video of low quality where heart structures are poorly visible, which can be caused, for instance, by adipose tissue thicker than normal between the probe and the patient's heart. Without any clearly detectable abnormalities, the network classified an RHD positive as negative, probably due to the low quality of the image. In Figures 3.7c and 3.7d the images are clear. Figure 3.7c shows the absence

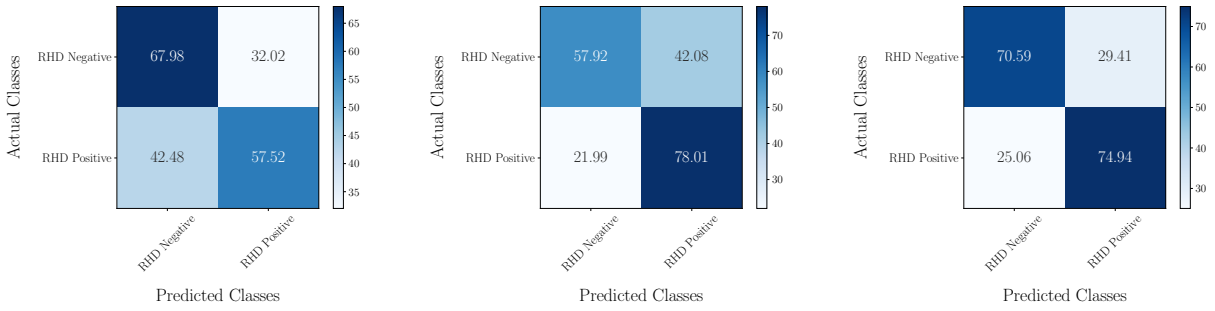


Figure 3.6: **Resulting confusion matrices for each method on Rheumatic Heart Disease classification of echocardiographic exams. (a)** VGG16 with Majority Vote **(b)** C3D with Majority Vote **(c)** C3D with Meta-Classifier.

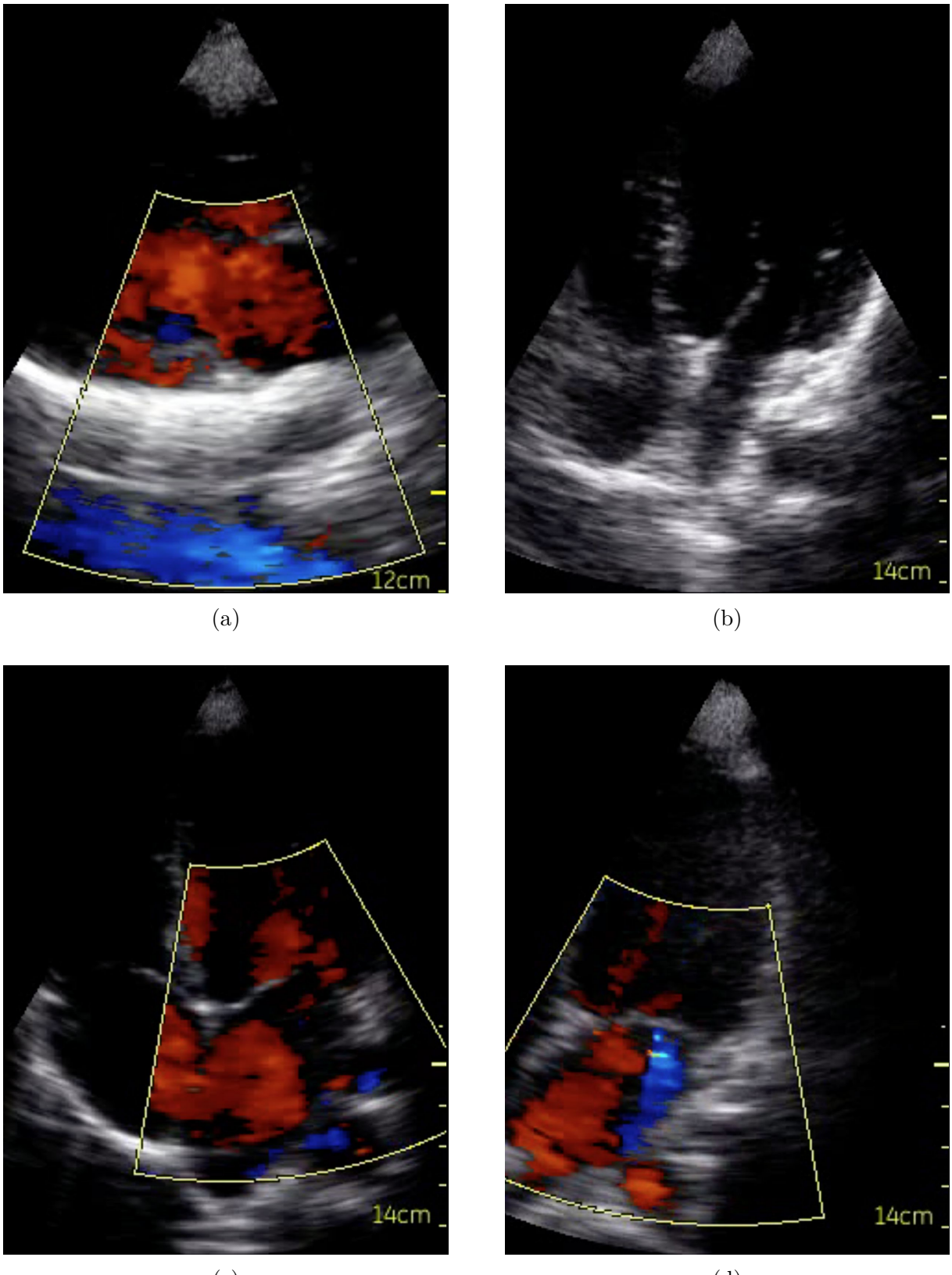
of mitral regurgitation during systole to represent the lack of abnormalities in the video, which led the model to correctly predict the exam as RHD negative. In Figure3.7d we can observe, also during systole, the presence of mitral regurgitation as the blue Doppler jet, which is one of the main factors for the detection of RHD, and therefore probably led the model to classify the video as RHD positive.

These examples show that the quality of images directly affects the performance of the model. However, as the main motivation for this work is to process the images as they come, a preprocessing step to remove this type of noise from the dataset can greatly improve the model's performance.

3.4 Summary

Throughout this chapter, we have described the first published method for automatic RHD diagnosis in conventional echocardiograms. The method's ability to encode temporal information was shown to be significantly more effective for the task at hand than previous methods from the literature, suggesting that our hypothesis was correct. We also show that the method is substantially better at identifying RHD at later stages of the disease, as opposed to the latent (Borderline) disease. The properties and effectiveness of the supervised meta-classifier for the final diagnosis of the exam are also discussed. Finally, we exemplify instances of the dataset to which the model issued a diagnosis with high confidence, in the process highlighting some of the issues present in our data.

In the next chapter, another method is proposed, this time providing multiple layers of interpretability while using medical annotations as auxiliary classification tasks to boost the accuracy in the main tasks of diagnosing RHD.



(a) (b)

(c) (d)

Figure 3.7: **Examples of frames extracted from 4 videos where the model made the predictions with high confidence.** Videos are from different exams, and we consider their predictions when in the test set. **(a)** RHD negative misclassified as RHD positive **(b)** RHD positive misclassified as RHD negative **(c)** RHD negative correctly classified **(d)** RHD positive correctly classified.

Chapter 4

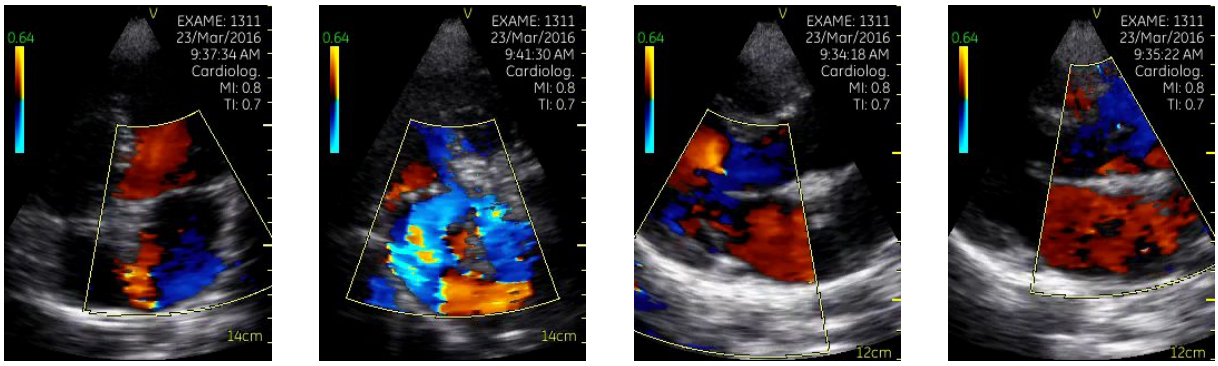
Interpretable Unsupervised Diagnosis with Multi-Task Learning

The confidence of physicians and patients toward automatic diagnosis plays an important role in the widespread adoption of computer-aided diagnosis in the real world. Therefore, having a method that provides interpretability of its predictions is crucial. Also, the scarcity of echo data from screenings shown in the previous chapter is something that will only change in the long term, so extracting more information relevant for diagnosis from the current dataset seems imperative.

The current chapter describes our final method for the automatic diagnosis of RHD in echocardiograms. We present a two-stream convolutional neural network in a multi-task learning setup that uses a 2D CNN as feature extractor but can nonetheless incorporate temporal information in the prediction through attention mechanisms. The method leverages labels of functional abnormalities of the heart as auxiliary tasks to increase its generalization ability. Furthermore, we propose an unsupervised aggregation strategy centered around detecting out-of-distribution videos as noisy instances, ultimately removing them from the final diagnosis process. The method is also able to provide interpretable information about its decision-making process on multiple levels.

4.1 Dataset

Our dataset is a subset of the one used in Chapter 3, and was composed by filtering for exams that contain annotations for Rheumatic Heart Disease (RHD) features related to Doppler videos. The data is comprised of 5,526 echocardiographic videos with resolution 320×240 pixels, taken with Vscan Extend™ devices. The videos correspond to 538 complete exams of unique patients (269 RHD negative and 269 RHD positive exams). The data was acquired by trained technicians only as part of the PROVAR screening program [76], as exams from the Uganda screening program [11] were lacking the anno-



(a) (b) (c) (d)

Figure 4.1: **Examples of frames sampled from four different viewpoints of a single exam.**

tations. Due to the extremely low number of samples of Definite RHD (only 32 exams in total), we once again aggregate both Borderline and Definite diagnosis into the same final label, RHD Positive.

As part of RHD's diagnosis procedure following the World Heart Federation's guidelines [88], doctors annotate for each exam different echocardiographic features that are directly related to morphological (structural) and functional abnormalities in the patient's heart valves. As we only use Doppler videos, the only relevant annotations will be those related to the function of the valves, i.e., the presence and nature of regurgitation and stenosis. Our dataset contains seven additional labels that are used as auxiliary binary classification tasks for our multi-task learning setup, namely: Mitral Stenosis, Mitral Regurgitation, Mitral Regurgitation $>$ 1.5cm, Mitral Regurgitation $>$ 2cm, Aortic Stenosis, Aortic Regurgitation, and Aortic Regurgitation $>$ 1cm.

The average age of patients was 13.1 (std = 3.1) years, from which 59.85% were female. Each exam contains, on average, 10.2 (std = 3.15) color Doppler echocardiograms videos, which visually depict the flow of blood through the heart's chambers and valves. Figure 4.1 shows frames extracted from videos of our dataset, which may depict one of four viewpoints: **(a)** Apical 4 Chambers with Doppler; **(b)** Apical 5 Chambers with Doppler; **(c)** Parasternal Long Axis with Doppler on the Mitral Valve Level; **(d)** Parasternal Long Axis with Doppler on the Aortic Valve Level.

4.2 Methodology

This section describes a new method for classifying RHD in an echocardiographic exam composed of a set of videos. Our method consists of two main steps: i) a multi-task

two-stream network trained for RHD detection and related sub-tasks with echocardiographic videos and ii) an unsupervised aggregation strategy that accounts for predictions across all videos to diagnose the exam. Figure 4.2 shows an overview of the two steps of our method.

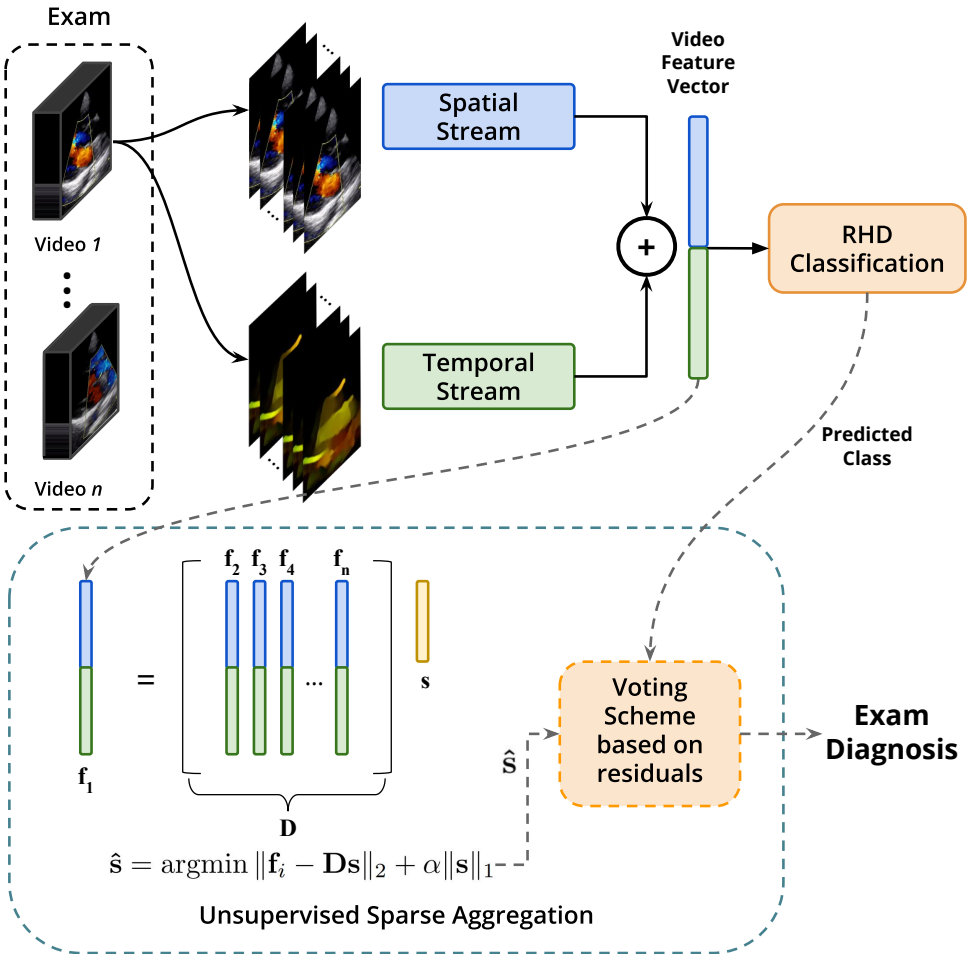


Figure 4.2: **Exam classification.** After computing the class score for each video using our multi-task two-stream network, we apply a sparse voting strategy that selects a few scores to determine the exam diagnosis.

4.2.1 Two-Stream Network

In diagnosing RHD in echocardiographic exams, functional (*e.g.*, regurgitation or stenosis) and structural (*e.g.*, restricted valve movement) abnormalities may indicate the disease's presence. These abnormalities can be identified in an echocardiographic video respectively via blood flow (color Doppler) and morphological analysis of the valves' movement. Even though no single B-Mode echocardiograms are used in the experiments de-

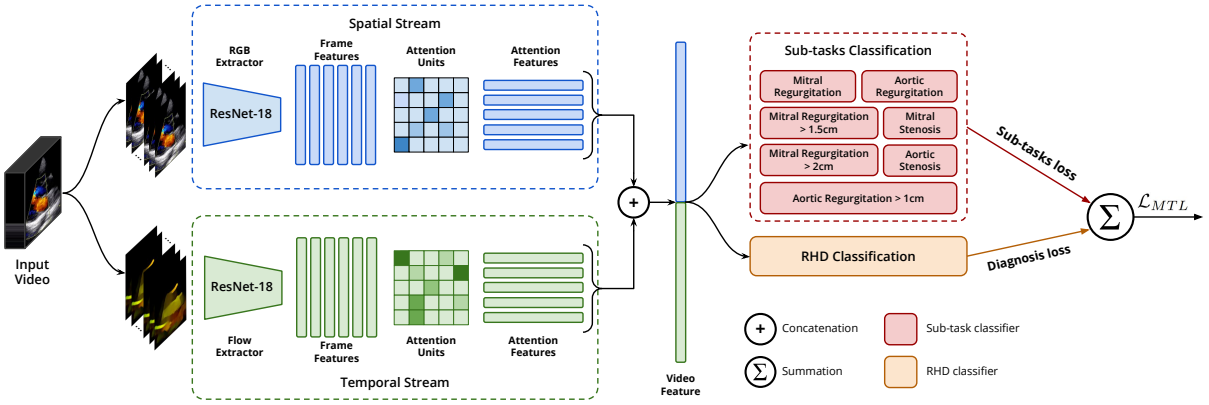


Figure 4.3: **Two-Stream Multi-Task Network training.** The networks are trained in a multi-task regime processing both spatial and temporal information from echocardiographic videos. Attention units are applied to weigh important frames, where morphological features and the blood flow are present.

scribed in this chapter, structural changes on the heart may still be visible outside the Doppler region of interest or when blood flow is small or absent, as color Doppler information is superimposed on B-Mode echo images. Our architecture simultaneously extracts relevant spatial and temporal features from visual data using two streams: the spatial feature extractor and the temporal feature extractor. Figure 4.3 depicts the architecture of our model and the training process.

Spatial Feature Extraction. In the spatial feature extractor stream, we fine-tune a ResNet-18 [51] pre-trained in the ImageNet dataset [30]. For each frame of our visual input, this stream extracts a feature vector of size 512 from the adaptive average pooling layer. Since our network receives as input RGB videos with 16 frames, we add duplicates of the first and last frames for videos shorter than 16 frames until the target length is achieved.

Temporal Feature Extraction. Identifying crucial features in RHD classification is directly correlated to the spatio-temporal observation of some heart structures. Thus, we incorporate the temporal relationship between consecutive frames and the structures represented in them into our model. We feed our second stream – the temporal feature extractor – with the optical flow generated by a frozen FlowNet 2.0 [55] for each of the 15 sequential pairs of frames present in a video, represented with the flow field color coding described by [55]. Similar to the spatial feature extractor, a pretrained ResNet-18 model is fine-tuned to extract the flow features.

Attention Units. Abnormalities appear in different moments of the heart's dynamic. Thus, these abnormalities may be detectable using a subset of frames of an echocardiogram. Aiming to weaken the influence of non-relevant or noisy frames from the video,

we use a set of attention units to aggregate the feature vectors of frames into the final representation.

Let X be a matrix in which columns are the feature vectors extracted from one of the network streams, *i.e.*, the spatial or temporal stream. We aggregate these feature vectors into the final representation g. For instance, let X_s be the matrix with feature vectors computed by the spatial stream network. We rely on an attention mechanism that learns the importance of each frame by computing the weights a_w. The weights are applied to X_s and generate the representation $a_f = X_s a_w$. Each attention unit is implemented as a single fully-connected layer with a softmax activation function as follows:

$$a_w = \text{softmax}(w^T X_s + b),$$

where w and b are the learnable attention weights.

While each attention unit focuses on a particular set of related frames, there can be multiple relevant parts that together describe the situation portrayed in the video. Therefore, we use the attention clusters approach proposed by [64], where, in order to efficiently lead different units to generate different weight distributions, a shifting operation is added to each attention unit. This operation is performed by applying a linear transformation to the original a_f, followed by a cluster-level ℓ_2 normalization, generating the shifted representation $\hat{a}_f$, defined as:

$$\hat{a}_f = \frac{c_1 X_s a_w + c_2}{\sqrt{N}\|c_1 X_s a_w + c_2\|_2},$$

where N is the number of attention units composing our attention cluster and both c_1 and c_2 are learnable scalars.

The final representation g_s of X_s is created by concatenating the representation $\hat{a}_f^i$ of each attention unit i: $g_s = \left[\hat{a}_f^1, \hat{a}_f^2, \cdots, \hat{a}_f^n\right]$, where n is the number of attention units composing our attention cluster.

The same procedure is simultaneously executed for the temporal stream network computing another set of weights w, b, and the final representation g_t.

4.2.2 Multi-Task Learning

In our multi-task learning approach, we leverage the additional labels described in Section 4.1 as auxiliary sub-tasks in our training procedure. The information contained in these 7 sub-tasks is used as a learning bias during training, helping the model with generalization and, therefore, increasing the final diagnosis's accuracy.

For each task i, we input the features extracted from the attention clusters in both streams into a fully-connected classifier. We use the binary cross-entropy loss as each classifier's loss $\mathcal{L}_i$:

$$\mathcal{L}_i(y, \hat{y}) = y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \log(\hat{y}_i)),$$

where y_i and $\hat{y}_i$ are the corresponding binary label and predicted probability for task i, respectively.

Let N be the number of tasks, including the RHD diagnosis. To learn representations that can be used in multiple tasks and enhance our generalization, our architecture minimizes the aggregated loss defined as:

$$\mathcal{L}_{MTL} = -\frac{1}{N} \sum_{i=0}^{N} \mathcal{L}_i.$$

4.2.3 Unsupervised Sparse Aggregation

Our method's ultimate goal is to diagnose an exam composed of a set of echocardiogram videos. In general, videos are not captured by specialized cardiologists, and most patients are children, who tend to be less cooperative during the procedure. These unfavorable conditions during the acquisition process lead to quality issues in some videos. During exam diagnosis, experts can identify and ignore misrecorded samples. However, our method will receive all instances regardless, as explained in Section 3.1. Since the outputs of our multi-task method are related to each video, we propose a new aggregation strategy that takes into consideration the possibility of having out-of-distribution videos, *i.e.*, videos that have acquisition problems and may not focus entirely on the heart (*e.g.*, videos including parts of the abdomen).

Let $\mathcal{V} = \{v_0, \cdots, v_{n-1}\}$ be an exam composed of n videos and $\mathbf{f}_i \in \mathbb{R}^m$ be the i-th feature vector extracted by our multi-task network from the i-th video. For each feature vector $\mathbf{f}_i$, we create the dictionary $\mathbf{D} \in \mathbb{R}^{m \times (n-1)}$ with columns composed of the remaining feature vectors $\mathbf{f}_j$, where $j \neq i$. Assuming that only a few videos contains good data, we solve a sparse code problem to represent the feature vector $\mathbf{f}_i$ using as few columns as possible from $\mathbf{D}$, *i.e.*, after creating the dictionary, for each feature vector $\mathbf{f}_i$, we solve the optimization:

$$\hat{\mathbf{s}} =_{\mathbf{s}} \|\mathbf{f}_i - \mathbf{D}\mathbf{s}\|_2 + \alpha \|\mathbf{s}\|_1,$$

where α is the penalty applied to the ℓ_1 norm and $\hat{\mathbf{s}}$ is the sparse representation of $\mathbf{f}_i$.

Each feature vector $\mathbf{f}_i$ votes twice. Firstly, the number of votes for class c, inferred by the network, is incremented by

$$\frac{1}{\|\mathbf{f}_i - \mathbf{D}\hat{\mathbf{s}}\|_2^2 \times (1 - SCI) + \epsilon},$$

where ϵ prevents zero division and SCI is the sparse concentration index of $\hat{\mathbf{s}}$ [122]. When the coefficients of the sparse vector $\hat{\mathbf{s}}$ are spread over the negative and positive classes, then $SCI = 0$, and when the coefficients are concentrated, $SCI = 1$.

Secondly, given the sparse representation of the i-th feature vector, according to the classification strategy used by [122], we compute, for both negative and positive classes, the characteristic functions $\delta_+ : \mathbb{R}^n \to \mathbb{R}^n$ and $\delta_- : \mathbb{R}^n \to \mathbb{R}^n$ that select the coefficients from $\hat{\mathbf{s}}$ associated with each class. Then, we encode the vectors $\delta_+(\hat{\mathbf{s}})$ and $\delta_-(\hat{\mathbf{s}})$ using matrix $\mathbf{D}$ and select the class which transformed vector is the closest to $\hat{\mathbf{s}}$. After selecting which class to vote, we increase the number of votes by inverse the squared error between the sparse representation and the transformed vector. The final classification of the exam is given by the class with more votes after processing all feature vectors.

4.3 Experiments

4.3.1 Baselines, Hyperparameters and Implementation

We compare our method against two baselines. First, we evaluate [125] method. Their method works in three steps: first, a VGG-13 CNN [101] predicts the disease separately for 10 random frames of a video; second, it uses the confidence mean per frame to determine the overall confidence and predict the class of a video; and finally, it calculates the median confidences of all the videos in an exam to issue a final diagnosis for a patient. Our second baseline is the method presented in Chapter 3, which classifies videos directly through a C3D backbone, then generates exam-level predictions with their supervised meta-classifier aggregation strategy. This time, we chose to use VGG-13 instead of VGG-16 to be accurate with the implementation in [125] and also for a more fair comparison with the C3D in terms of method size (learnable parameters). For a clearer comparison, from now on, we will refer to Chapter 3's method as Martins *et al.*, as it is the foundational published work for automatic RHD diagnosis in conventional echocardiograms [72].

The VGG-13 model was pre-trained on the ImageNet dataset, while the C3D model was pre-trained on the Sports-1M [58] dataset. All methods were implemented

using Python 3.6 in PyTorch 1.2 [82], and executed in a machine with Intel(R) Xeon(R) E5-2620 CPU and an NVIDIA TITAN RTX GPU.

We selected our hyperparameters using the performance of the models in validation partitions. Our method was trained during 25 epochs using the SGD optimizer with a batch size of 8, a learning rate of $1e-2$, and a momentum of 0.9. Regarding the attention cluster size, 32 heads are used for both streams. Convolutional layers were kept frozen in the first 5 epochs for all models. The learning rate used for the VGG-13 was $1e-4$, while the learning rate for the C3D was $1e-3$. We used the Least-angle regression (LARS) [34] optimization with $\alpha = 2.0$ to estimate the sparse vectors $\hat{\mathbf{s}}$ in the sparse aggregation step.

4.3.2 Experimental Setup

We have performed a binary classification with the Borderline RHD and Definite RHD diagnosis grouped into a single class named RHD Positive. All information outside of the ultrasound beam was removed for de-identification purposes. The images were directly downsized to the input size of each method, normalized, and had their color channels centralized using the respective mean and standard deviation obtained in the training dataset at each iteration. We extracted the optical flow using FlowNet 2.0 [55].

We used a 10-fold cross-validation procedure in our experiments. For each split, we randomly divided the data into training, validation, and test partitions in an approximate 80:10:10 ratio and each video went through the validation and test partitions only once. We ensured that same-exam videos were always in the same partition, and splits were stratified according to the diagnosis class. For fair pairwise comparison, splits are fixed for all evaluated methods. The quantitative analysis uses the mean specificity, sensitivity, and accuracy score metrics for the test partitions.

4.3.3 Results

We evaluate the efficacy of RHD classification in video and exam levels. While video-level predictions are made from a single video, exam-level predictions result from aggregating the video-level predictions using strategies such as majority vote and the proposed sparse aggregation.

Table 4.1 reports the mean specificity, sensitivity, and accuracy scores for RHD

Table 4.1: **Comparison with baselines.** Average specificity, sensitivity, and accuracy for RHD classification using 10-fold cross-validation. Best values in bold according to a 95% confidence Wilcoxon signed-rank test. Sizes represent the number of learnable parameters in each of the methods.

Method	Size	Video			Exam		
		Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
[125]	129M	34.33 ± 9.02	**76.10** ± 6.92	56.87 ± 3.42	31.29 ± 10.55	**83.30** ± 7.26	57.26 ± 3.93
Supervised Temporal-Aware	130M	58.57 ± 10.51	58.00 ± 10.59	58.20 ± 2.70	64.84 ± 7.03	65.14 ± 10.84	64.88 ± 5.22
Unsupervised Multi-Task	23M	**65.87** ± 7.81	63.83 ± 8.46	**64.70** ± 2.17	**71.71** ± 10.90	70.70 ± 9.62	**71.18** ± 3.10

classification in the test partitions, both in the video and exam levels. To ensure statistical significance in comparing the methods' performance, we used a Wilcoxon-signed rank test with a confidence interval of 95%. We draw the following observations. Regarding accuracy, the temporal-aware methods, namely Martins *et al.* and ours, outperformed the frame-based predictor from Zhang *et al.* in all aggregation levels. Our method performs significantly better than Martins *et al.*, the current state-of-the-art for RHD classification while containing 5× fewer parameters. The same pattern can be observed in the specificity results.

Although the method of Zhang *et al.* achieved the best sensitivity results, it is worth noting the significant discrepancy between its specificity and sensitivity, which reaches values of 98.01% in sensitivity versus 8% specificity at the video-level for one of the splits of the cross-validation procedure. This discrepancy is even more prominent when the results are aggregated to obtain the exam-level predictions. We argue that this happens because the method predicts frame by frame without any temporal correlation, and, thus, its generalization ability is handicapped due to the nature of RHD classification, which is strongly influenced by temporal cues such as the blood flow. There is a clear bias towards the positive diagnosis in this classifier, which ends up generating a lot of false positives in the process. After aggregation, the video-level bias is propagated to the exam-level diagnosis, increasing the bias even further.

For sanity checking purposes, we repeat the evaluation of results done in Section 3.3.3 for the subclasses that compose the RHD Positive class: Borderline RHD and Definite RHD. The overall sensitivity of 70.70% (± 9.62) is broken into 69.59% (± 10.12) for the Borderline subclass and 85.50% (± 9.31) for the Definite subclass, repeating the pattern seen previously and corroborating that our model is learning the correct features for RHD diagnosis, which become more noticeable in more severe stages of the disease.

4.3.4 Ablation Study

To verify the contribution of the main components of our methodology, namely Sparse Aggregation, Multi-Task Learning, Two-Stream Network, and Attention Units to the success of our approach, we progressively remove each component while executing the same training procedure and evaluate the impact on the exam classification. Table 4.2 shows the results after removing each of the following components:

Aggregation Strategies. We compare our unsupervised Sparse Voting aggregation (TSM_{SV}) with the supervised Meta-Classifier presented in [72] (TSM_{MC}), and the majority vote strategy (TSM_{MV}). We argue that the sparse formulation identifies out-of-distribution videos that should not be considered valid votes without depending on a large exam set to counterbalance biases acquired during model training due to its non-supervised nature. The majority vote strategy is outperformed by its counterparts, presenting an accuracy score 1.68% lower when compared to the (TSM_{SV}), evidencing the importance of more sophisticated aggregation strategies for the task at hand. However, while TSM_{SV} is significantly better than TSM_{MV}, given a 95% confidence Wilcoxon signed-rank test, the same is not true when comparing the latter with TSM_{MC}, reinforcing the importance of the proposed aggregation strategy. These results indicate that supervised aggregation strategies can struggle in more realistic scenarios where fewer data tend to be available. We can also observe that the specificity and sensitivity are more balanced in the sparse aggregation results. This balance is desirable in our context. Finally, the unsupervised method can provide a new layer of interpretability to exam level classification by indicating which videos contributed the most to the voting process that resulted in the final diagnosis.

Table 4.2: **Ablation study.** Effects on RHD classification for different components of our method: Spatial stream only (OS) with a global average strategy (OS_{GA}) and with attention units (OS_{att}); Two-Stream only (TS); multi-task approaches with Majority vote (TSM_{MV}), Meta-Classifier (TSM_{MC}), and Sparse voting (TSM_{SV}). All results are for exam-level using 10-fold cross-validation. Where no aggregation strategy is explicit, Majority Vote was used.

Method	Specificity	Sensitivity	Accuracy
OS_{GA}	51.67 ± 8.63	71.84 ± 11.52	61.72 ± 4.36
OS_{att}	54.98 ± 14.86	76.64 ± 7.73	65.64 ± 5.28
TS	57.01 ± 15.16	75.13 ± 10.20	65.97 ± 5.30
TSM_{MV}	72.83 ± 10.25	66.24 ± 11.87	69.50 ± 2.93
TSM_{MC}	66.11 ± 9.46	74.76 ± 7.84	70.43 ± 4.51
TSM_{SV}	71.71 ± 10.90	70.70 ± 9.62	71.18 ± 3.10

Multi-Task Learning. To ascertain that our multi-task approach (TSM_{SV}) improves the generalization capacity of our method, we contrast it with a model that only predicts the RHD diagnosis (TS). The addition of auxiliary tasks regarding the classification of the heart's functional characteristics resulted in 3.53% net improvement in the accuracy performance for our main task. The expressive reduction of the confidence interval size for the accuracy, primarily due to significant improvements in specificity, corroborates the importance of our multi-task setup. When experimenting with annotations for morphological features (along with B-Mode echos), the network presented convergence problems. This makes complete sense, given that the prevalence of morphological features in subclinical disease is very low [88]. Therefore additional information was very poor and not able to improve generalization ability. Training the network with auxiliary tasks related to both functional and morphological features was also less effective, as functional abnormalities are not observable in B-Mode echos.

Two-Stream Network. As stated, the addition of a temporal stream is prompted by the observed correlation between spatial-temporal observation of some heart structures with the RHD diagnosis. For this analysis, we removed the flow stream from our single-task architecture (TS) and kept the remaining settings (OS_{att}). Although we only improved the overall accuracy marginally, the results show that the classifier was more balanced in its predictions regarding specificity and sensitivity. In Section 4.2.1 we hypothesize that, even without B-Mode echos, morphological features might still be present in the videos used. However, this result shows that for the current experiment, this was probably not the case. Nonetheless, it is important to remember that this method may be used in a different experimental setup, preventing the complete invalidation of the proposed temporal stream for now.

Attention Units. Lastly, we replace the attention units with a straightforward global average strategy (OS_{GA}), which can be interpreted as a degenerate form of attention [64]. This experiment allows us to assess the contribution of the learned attention weighting the frames when generating the video feature vector. As stated, not all frames in echocardiograms contain information useful for diagnosing the disease. Frames that do not contain blood flow, for instance, cannot show signs of any functional abnormalities, such as mitral regurgitation, which can hint at the presence of RHD. We assume that distributing equal weights among frames would result in lower predictive performance. Results indicate that our hypothesis holds, since the accuracy increased by 3.92% points in total, with improvements in both specificity and sensitivity compared to the global average operation to generate the feature vectors classified upon by the fully connected layer.

4.3.5 Interpretability

Physicians' and patients' confidence toward automatic diagnosis plays an important role in the widespread adoption of computer-aided diagnosis applications [111, 106]. Our method is able to boost users' confidence by generating two visualizations that work together, bringing a sense of interpretability to the results. First, our method provides the importance of each frame given by the matrix of weights embedded within each stream's attention units. The final values assigned to each frame are generated by taking the mean of the weight vectors outputted by each of the clusters' attention units.

Additionally, after identifying which frames are more relevant, we can also generate the Class Activation Maps (CAMs) for the frames of interest to check which image regions were more important to generate the prediction. The CAMs are generated using the Score-CAM [113] method after a forward pass in the last layer of our ResNet-18 extractors. We chose to use Score-CAM because it achieves better visual performance and fairness for interpreting the decision-making process [113].

Figure 4.4 shows a way of interpreting the model's decision process for a single video. In this case, information that stands out in the Spatial Stream will mainly be related to blood flow abnormalities that may indicate pathological valve regurgitation or stenosis. At the same time, the Temporal Stream interpretation possibly highlights structural movement of the heart that was important for RHD diagnosis. Further information for video-level interpretation could be provided by the predicted labels for the sub-tasks in each instance.

Concerning exam-level interpretation, the proposed unsupervised aggregation strategy enables the identification of the videos in the exam that contributed significantly to the final agreed diagnosis, thus possibly reducing the number of samples that need to be inspected. This identification can be made by assessing each video participation during the voting procedure described in Section 4.2.3.

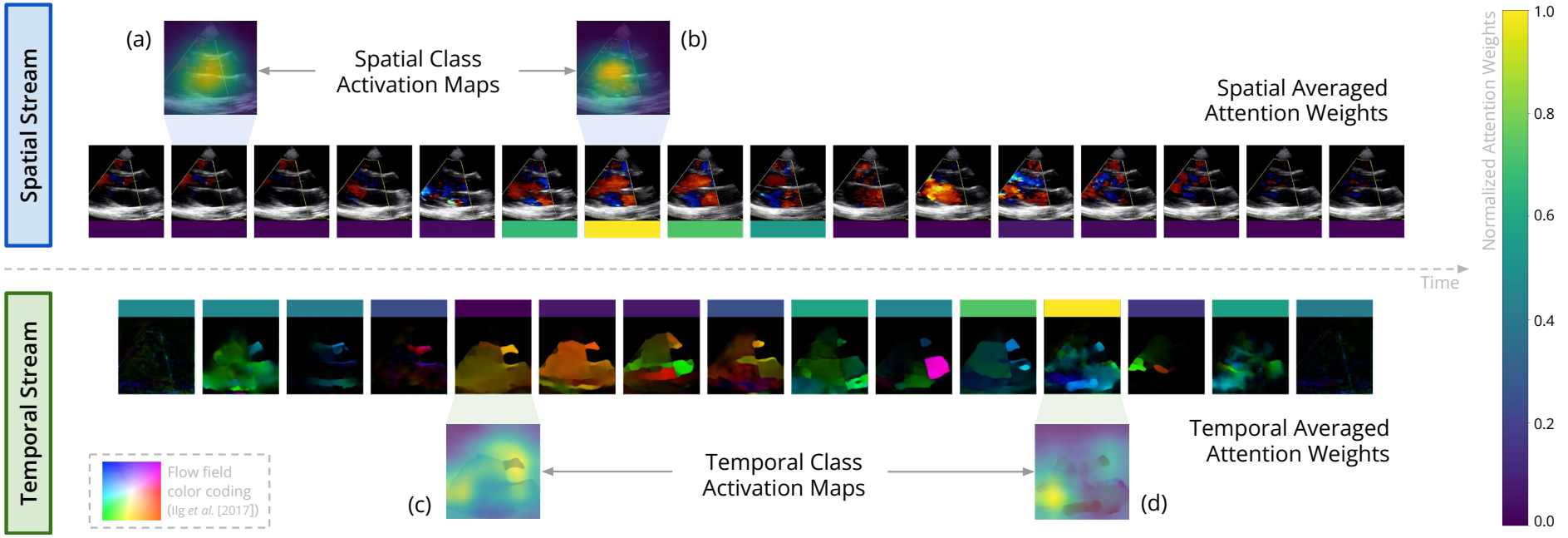


Figure 4.4: **Video-level interpretability of our method.** For each stream, two frames are highlighted with their respective ScoreCAM visualization. In the spatial stream, both most **(a)** and least **(b)** attended frames have similar activation maps, emphasizing that the actual region of interest (where blood flow is detected) is contributing the most to the model's prediction. However, **(b)** represents a case of mitral regurgitation at its peak (blue blood flow in the original frame), which an expert later measured to be 2.3cm long, while **(a)** contains almost no blood flow. Regarding the temporal stream, the first highlighted frame **(c)** is an example of the method's ability to pay less attention to frames in which the most activated area is actually outside the region of interest while attending to more relevant structural movement as shown in frame **(d)**.

4.3.6 Practical Application

Concerning the diagnosis, better sensitivity would be the preferred metric because misdiagnosing a patient's disease in its initial stages would let it progress and create a burden for their quality of life and the government's public health system. However, the shortage of expert personnel and equipment combined with a biased classifier reporting patients diagnosed with RHD, even in large urban areas, is a limiting factor to how many patients can be referred to cardiac centers from screenings in remote areas. Also, from the perspective of patients, positive results have been shown to cause anxiety [50] and decreased parental [16] and child quality of life [16, 115], which are the most significant causes of harm in RHD screenings. There is a lack of broader studies that weigh the benefits over the harms of screenings [32], making assumptions about preferred metrics more difficult. Therefore, we argue that a balanced specificity and sensitivity may be the most prudent until further studies are conducted.

There are two studies that explore non-expert diagnoses that not only have a setup similar to the one used in this experiment but also were conducted with the same data that compose the dataset described in Section 3.1, coming from Uganda [84] and Brazil [12]. The main difference consists in their use of abbreviated diagnostic criteria to issue a positive diagnosis, named screen positive. The criteria are simply the identification of mitral regurgitation greater than 1.5 cm or any aortic regurgitation. Therefore, screen positive diagnosis seems to be a much easier task than RHD Positive diagnosis. In Table 4.3 we present a comparison of the Specificity and Sensitivity of the methods described in this dissertation with the performance of non-experts in the relevant studies.

[84] trained two nurses with 4 hours of theoretical material and 2 days of echocardiographic practice. They report a specificity of 78.8% (95% CI, 76.0%-81.4%) and a sensitivity of 74.4% (95% CI, 58.8%-86.5%), which was judged reasonable by the authors. [12] trained 2 nurses, 2 medical students and 2 biotechnicians with a much more thorough 3-week computer-based interpretation course and weekly hurdle assessments. They report more expressive metrics, with a specificity of 85% (95% CI, $82\% - 87\%$) and a sensitivity of 83% (95% CI, $76\% - 89\%$), performances comparable to previous handheld screening

Table 4.3: **Comparison with non-experts.**

Source	Specificity	Sensitivity
Unsupervised Multi-Task Subset	71.7 (60.8, 82.6)	70.7 (61.1, 80.3)
Supervised Temporal-Aware Full	70.6 (66.5, 74.6)	74.9 (70.1, 79.7)
Ploutz et al. [2016]	78.8 (76.0, 81.4)	74.4 (58.8, 86.5)
Beaton et al. [2016]	85.0 (82.0, 87.0)	83.0 (76.0, 89.0)

performance by both expert and non-expert users [12]. The performance of our method came close to the one reported by the study with less intensive training of non-experts, [84]. However, experiments in the automatic diagnosis of screen positive RHD should be performed for a fairer comparison.

Given that a computer-aided diagnosis system would represent a significant step in overcoming financial and workforce barriers that limit widespread RHD screening due to its diagnosis speed and low-maintenance costs, we argue that the efficacy obtained in this work would be sufficient for practical use. The development of applications for real-world adoption of this technology is discussed more in depth in the next chapter.

4.4 Summary

Throughout this chapter, we present our final method for the automatic diagnosis of RHD in echocardiograms. It comprises a two-stream attention-based 2D CNN within a multi-task learning setup and an unsupervised sparse voting strategy for exam diagnosis. Our new method is not only able to significantly outperform other baselines with an accuracy of 71.18% but is also able to provide consistent information about its decision-making process in multiple levels, mainly as temporal (relevant frames in the video) and spatial (relevant structures in a frame) visualizations, which is depicted and discussed. An ablation study is performed to understand the contribution of each component in the method, and comparison with non-expert human performance is also drawn.

The next chapter concludes this dissertation, wrapping up our contributions and outlining multiple directions for future work.

Chapter 5

Conclusion and Future Work

This work lays the foundations for automatically diagnosing RHD in conventional echocardiographic exams through machine learning algorithms.

In Chapter 3, we test our hypothesis that a temporal-aware method would perform better than the current literature for disease identification in echocardiograms by proposing the use of a 3D convolutional neural network (CNN), C3D [108], for individual video classification. We also propose a more sophisticated aggregation strategy to issue a whole exam diagnosis, which is supervised and based on the confidence distribution for the video predictions of the previous classifier. Experiments show that the temporal-aware method and also the supervised aggregation strategy are significantly better at the task of RHD identification. The work described in this chapter is the foundational publication for automatic RHD diagnosis in conventional echocardiograms [72].

Chapter 4 presents a two-stream attention-based convolutional neural network that leverages annotations of sub-tasks naturally created by experts during the diagnosis procedure to improve RHD classification accuracy significantly. The complete method is also composed of an unsupervised sparse voting strategy that aggregates video predictions into exam-level diagnosis by accounting for out-of-distribution samples. Our approach significantly outperforms baseline methods, achieving state-of-the-art performance while providing strong interpretability, facilitating the adoption of the method in clinical decision-making. The unsupervised aggregation strategy was significantly better than the previously implemented supervised meta-classifier in the experimental setup described.

RHD diagnosis using conventional echocardiograms is a challenging problem due to the extensive diagnosis guidelines that need to be used along with different types of data (multiple echo modes and viewpoints). Data is also scarce and populated with noise from different sources. Moreover, the existing literature is minimal, and no previous related works used methods suitable for the task approached in this study. Nonetheless, automatic diagnosis of echo-detected RHD seems feasible and, with further research, has the potential to substantially reduce the workload on cardiologists and experts, enabling the implementation of more widespread screening programs that can reduce the disease burden in the underdeveloped world. More than the point-of-care and telemedicine diagnosis of RHD, the proposed system, embedded in screening devices or made available as a

cloud-based application, also has the potential to allow low-cost identification of patients at higher risk for other valvulopathies and cardiovascular diseases. Future works in RHD or similar diseases will also greatly benefit from the framework for interpretable video and exam classification described in Chapters 3 and 4. Even though we work with images from handheld devices, the same methods can be used with echocardiograms obtained with different types of machines.

5.1 Future Work

Immediate improvements to the current setup can be made:

- **Train our method for the task of screen positive RHD diagnosis.** As described in Section 4.3.6, studies about the effectiveness of task-shifting RHD diagnosis to trained non-experts mostly use abbreviated diagnostic criteria, which is possibly an easier classification task. As the dataset used in Chapter 4 have the annotations for functional features of RHD, we can reproduce the labels for screen positive diagnosis as described in [84] and [12] and train our method with these labels, instead of the more broad RHD diagnosis.
- **Implement more powerful feature extractors to boost accuracy.** Powerful new methods, such as the most recent visual transformers [63], can not only improve classification accuracy but also provide better interpretability through more detailed attention maps [22].
- **Actively remove noise instances from exams before issuing a diagnosis.** It is clear at this point that some exams have more than one instance per viewpoint and that probably some of them will have quality issues, as discussed for Figure 3.7. Removing them *a priori* will possibly boost classification accuracy, as less noise will be considered during the aggregation step that issues the final diagnosis. The unsupervised aggregation strategy presented in Chapter 4 tries to mitigate this problem, but we could also, for instance, train an additional supervised classifier for viewpoint identification and remove instances that fall below a confidence threshold when going through it.

One of the main problems we face in this work is the low amount of training data. In our largest experiment, presented in Chapter 3, the classifiers see at most 365 exams with a positive diagnosis for RHD per partition during training, from which only 26 would

be from people with the Definite diagnosis, thus having more expressive visual features. Acquiring new exams from the same or similar screening programs would not only be important for the assessment of the robustness and generalization of the methodology proposed but would open up new experimental possibilities:

- **Address the RHD diagnosis problem considering all classes.** We currently do not work with all classes due to data scarcity for the definite class, as screenings are directed towards subclinical (borderline) cases of the disease. However, issuing separate diagnoses would help the referral process for immediate healthcare assistance.

- **Train separate classifiers for each type of viewpoint.** With more data, training separate classifiers for each viewpoint may also become feasible, them allowing for the creation of a rule-based aggregation strategy adapted from diagnosis guidelines established by the World Heart Federation [88]. We performed viewpoint classification using pre-trained networks for the task and subsequently trained our method for RHD diagnosis with specific viewpoints from our current dataset. The first step showed good results, but in the latter, networks would not converge due to the very reduced size of datasets after splitting by viewpoint. Also, as some viewpoints may be focused in specific structures, a subset of the sub-tasks may be more suitable for a multi-task learning setup within the separate classifiers.

- **New instance generation using Generative Adversarial Networks (GANs).** With enough new samples, Generative Adversarial Networks could be used to generate even more instances, similar to what is shown in [2] or [37], which could be used for training and improving our classifiers.

Finally, regarding the adoption of our work in the real world, there are two main approaches:

- **Cloud-based application for remote diagnosis.** Creating an application in the cloud would remove the need for centralized infrastructure while also allowing the upload of just acquired exams for automatic diagnosis while patients are still at the point-of-care facility (if internet connection is available), thus reducing the uncertainty of referrals done by trained technicians. Such an application could also be used later to reduce the workload on expert cardiologists, as the system could predict the diagnosis for an entire batch of exams, with the expert only reviewing results in which the system's confidence was not very high. Using an established cloud platform, e.g., Amazon Web Services, would simplify the process, and costs could be greatly reduced by joining their Nonprofits & NGO program [6].

- **Diagnosis system embedded into portable devices.** Some screening locations may not have access to the internet, however, automated diagnosis methods could be used if they were embedded into portable devices, such as the handheld echo devices themselves. Newer handheld devices are even using standalone probes that connect to smartphones, as depicted in Figure 5.1, reducing application development barriers. In order to create embedded diagnosis systems, the application size and computation required for a prediction need to be reduced as much as possible. This can be done by implementing the trained deep neural networks using TensorFlow Lite [47] instead of PyTorch, which generates a much smaller and optimized model. We can also reduce the trained model size even further by using pruning techniques such as the Lottery Ticket Hypothesis [41]. Such models could even be used to detect inconsistencies in the capture and instruct technicians to redo the procedure.

Figure 5.1: **Standalone ultrasound probe that is connected with a smartphone.** This is a Vscan Air™ wireless handheld ultrasound device from GE Healthcare.

Bibliography

[1] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] A. Abdi, T. Tsang, and P. Abolmaesumi. Gan-enhanced conditional echocardiogram generation. *arXiv preprint arXiv:1911.02121*, 2019.

[3] G. Anabwani and P. Bonhoeffer. Prevalence of heart disease in school children in rural kenya using colour-flow echocardiography. *East African medical journal*, 73(4):215–217, 1996.

[4] A. Andermann, I. Blancquaert, S. Beauchamp, and V. Déry. Revisiting wilson and jungner in the genomic age: a review of screening criteria over the past 40 years. *Bulletin of the World Health Organization*, 86:317–319, 2008.

[5] A. Arguedas and E. Mohs. Prevention of rheumatic fever in costa rica. *The Journal of pediatrics*, 121(4):569–572, 1992.

[6] AWS. Aws for nonprofits and ngos. `https://aws.amazon.com/government-education/nonprofits`, 2021. Accessed: 2021-09-25.

[7] J. Bach, S. Chalons, A. Mosser, E. Forier, G. Elana, J. Jouanelle, S. Kayemba, D. Delbois, C. Sainte-Aimé, and C. Berchel. 10-year educational programme aimed at rheumatic fever in two french caribbean islands. *The Lancet*, 347(9002):644–648, 1996.

[8] S. Barnes et al. Echocardiographic screening of schoolchildren in american samoa to detect rheumatic heart disease: A feasibility study. *Pediatric Health, Medicine and Therapeutics*, 2:21–23, 2011.

[9] P. Barrett and E. Topol. To truly look inside. *Lancet (London, England)*, 387(10025):1268–1269, 2016.

[10] A. Beaton et al. The utility of handheld echocardiography for early diagnosis of rheumatic heart disease. *Journal of the American Society of Echocardiography*, 27(1):42–49, 2014.

[11] A. Beaton et al. The utility of handheld echocardiography for early rheumatic heart disease diagnosis: a field study. *European Heart Journal-Cardiovascular Imaging*, 16(5):475–482, 2015.

[12] A. Beaton et al. Efficacy of a standardized computer-based training curriculum to teach echocardiographic identification of rheumatic heart disease to nonexpert users. *The American journal of cardiology*, 117(11):1783–1789, 2016.

[13] J. Betts et al. Chapter 19 - the cardiovascular system: The heart. In *Anatomy and Physiology*. OpenStax, 2013.

[14] C. Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[15] H. Bleich. Computer-based consultation: electrolyte and acid-base disorders. *The American journal of medicine*, 53(3):285–291, 1972.

[16] T. Bradley-Hewitt et al. The impact of echocardiographic screening for rheumatic heart disease on patient quality of life. *The Journal of pediatrics*, 175:123–129, 2016.

[17] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[18] J. Carapetis et al. Acute rheumatic fever and rheumatic heart disease. *Nature Reviews Disease Primers*, 2(1):1–24, 2016.

[19] J. Carapetis, J. Parr, and T. Cherian. Standardization of epidemiologic protocols for surveillance of post-streptococcal sequelae: acute rheumatic fever, rheumatic heart disease and acute post-streptococcal glomerulonephritis. *National Institute of Allergy and Infectious Diseases*, 1, 2006.

[20] J. Carapetis, A. Steer, E. Mulholland, and M. Weber. The global burden of group a streptococcal diseases. *The Lancet infectious diseases*, 5(11):685–694, 2005.

[21] G. Carneiro, J. Nascimento, and A. Freitas. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. *IEEE Trans. Image Process.*, 21(3):968–982, March 2012.

[22] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[23] F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[24] K. Cios, K. Chen, and R. Langenderfer. Use of neural networks in detecting cardiac diseases from echocardiographic images. *IEEE Engineering in Medicine and Biology Magazine*, 9(3):58–60, 1990.

[25] P. Coffey, A. Ralph, and V. Krause. The role of social determinants of health in the risk and prevention of group a streptococcal infection, acute rheumatic fever

and rheumatic heart disease: a systematic review. *PLoS neglected tropical diseases*, 12(6):e0006577, 2018.

[26] M. Cohen et al. Racial and ethnic differences in the treatment of acute myocardial infarction. *Circulation*, 121(21):2294–301, 2010.

[27] S. Colquhoun et al. Pilot study of nurse-led rheumatic heart disease echocardiography screening in fiji–a novel approach in a resource-poor setting. *Cardiology in the Young*, 23(4):546–552, 2013.

[28] A. M. Davis, L. M. Vinci, T. M. Okwuosa, A. R. Chase, and E. S. Huang. Cardiovascular health disparities. *Medical Care Research and Review*, 64(5_suppl):29S–100S, 2007.

[29] J. De Fauw et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255, 2009.

[31] G. DiSciascio and A. Taranta. Rheumatic fever in children. *American heart journal*, 99(5):635–658, 1980.

[32] S. Dougherty, J. Carapetis, L. Zühlke, and N. Wilson. Acute rheumatic fever and rheumatic heart disease. Elsevier Health Sciences, 2020.

[33] P. S. Douglas et al. Accf/ase/aha/asnc/hfsa/hrs/scai/sccm/scct/scmr 2011 appropriate use criteria for echocardiography. *Journal of the American College of Cardiology*, 57(9):1126–1166, 2011.

[34] B. Efron et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[35] D. Engelman et al. Focused cardiac ultrasound screening for rheumatic heart disease by briefly trained health workers: a study of diagnostic accuracy. *The Lancet Global Health*, 4(6):e386–e394, 2016.

[36] D. Engelman, C. Watson, B. Remenyi, and A. Steer. Echocardiographic diagnosis of rheumatic heart disease. In *WiRED International Medical Education Programs*. WiRED International, 2015.

[37] Y. Engr, A. Lalande, J. Afilalo, and P.-M. Jodoin. Generative adversarial networks in cardiology. *Canadian Journal of Cardiology*, 2021.

[38] A. Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[39] A. Esteva et al. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[40] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[41] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[42] X. Gao, W. Li, M. Loomes, and L. Wang. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion*, 36:103–113, 2017.

[43] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[44] A. Ghorbani et al. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1):1–10, 2020.

[45] J. Godown et al. Handheld echocardiography versus auscultation for detection of rheumatic heart disease. *Pediatrics*, 135(4):e939–e944, 2015.

[46] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[47] Google. Tensorflow lite: Deploy machine learning models on mobile and iot devices. `https://www.tensorflow.org/lite`, 2021. Accessed: 2021-10-31.

[48] L. Gordis. The virtual disappearance of rheumatic fever in the united states: lessons in the rise and fall of disease. t. duckett jones memorial lecture. *Circulation*, 72(6):1155–1162, 1985.

[49] G. Gorry, J. Kassirer, A. Essig, and W. Schwartz. Decision analysis as the basis for computer-aided management of acute renal failure. *The American journal of medicine*, 55(4):473–484, 1973.

[50] J. Gurney, A. Chong, N. Culliford-Semmens, E. Tilton, N. Wilson, and D. Sarfati. The benefits and harms of rheumatic heart disease screening from the perspective of the screened population. *International journal of cardiology*, 221:734–740, 2016.

[51] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[52] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[53] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, 8, 2015.

[54] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[55] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. CVPR*, pages 2462–2470, 2017.

[56] S. James et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[57] M. Jowett, M. Brunal, G. Flores, and J. Cylus. Spending targets for health: no magic number. Technical report, World Health Organization, 2016.

[58] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, pages 1725–1732, 2014.

[59] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[60] T. Lamagni et al. The epidemiology of severe streptococcus pyogenes associated disease in europe. *Eurosurveillance*, 10(9):9–10, 2005.

[61] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[62] R. Ledley. Digital electronic computers in biomedical science. *Science*, 130(3384):1225–1234, 1959.

[63] Y. Liu et al. A survey of visual transformers. *arXiv preprint arXiv:2111.06091*, 2021.

[64] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proc. CVPR*, June 2018.

[65] A. Lu, E. Dehghan, G. Veni, M. Moradi, and T. Syeda-Mahmood. Detecting anomalies from echocardiography using multi-view regression of clinical measurements. In *15th Int. Symposium on Biomedical Imaging*, pages 1504–1508. IEEE, 2018.

[66] K. Macleod, P. Bright, A. Steer, J. Kim, D. Mabey, and T. Parks. Neglecting the neglected: the objective evidence of underfunding in rheumatic heart disease. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 113(5):287–290, 2019.

[67] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digital Medicine*, 1(1):6, 2018.

[68] A. Madani, J. R. Ong, A. Tibrewal, and M. R. K. Mofrad. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine*, 1(1):1–11, 2018.

[69] H. Majeed, S. Batnager, A. Yousof, F. Khuffash, and A. Yusuf. Acute rheumatic fever and the evolution of rheumatic heart disease: a prospective 12 year follow-up report. *Journal of clinical epidemiology*, 45(8):871–875, 1992.

[70] E. Marijon et al. Prevalence of rheumatic heart disease detected by echocardiographic screening. *New England Journal of Medicine*, 357(5):470–476, 2007.

[71] C. Martin-Isla et al. Image-based cardiac diagnosis with machine learning: A review. *Frontiers in Cardiovascular Medicine*, 7:1, 2020.

[72] J. Martins et al. Towards automatic diagnosis of rheumatic heart disease on echocardiographic exams through video-based deep learning. *Journal of the American Medical Informatics Association*, 2021. doi:10.1093/jamia/ocab061.

[73] M. Mirabel. Screening for rheumatic heart disease: evaluation of a focused cardiac ultrasound approach. *Circulation: Cardiovascular Imaging*, 8(1):e002324, 2015.

[74] C. Mondo et al. Presenting features of newly diagnosed rheumatic heart disease patients in mulago hospital: a pilot study. *Cardiovascular journal of Africa*, 24(2):28, 2013.

[75] M. Naghavi et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1151–1210, 2017.

[76] B. Nascimento et al. Echocardiographic prevalence of rheumatic heart disease in brazilian schoolchildren: Data from the provar study. *Int. journal of cardiology*, 219:439–445, 2016.

[77] B. Nascimento et al. Rheumatic heart disease echocardiographic screening: approaching practical and affordable solutions. *Heart*, 102(9):658–664, 2016.

[78] B. Nascimento et al. Comparison between different strategies of rheumatic heart disease echocardiographic screening in brazil: data from the provar (rheumatic valve disease screening program) study. *Journal of the American Heart Association*, 7(4):e008039, 2018.

[79] P. Nordet, R. Lopez, A. Duenas, and L. Sarmiento. Prevention and control of rheumatic fever and rheumatic heart disease: the cuban experience (1986-1996-2002): cardiovascular topic. *Cardiovascular journal of Africa*, 19(3):135–140, 2008.

[80] E. Okello et al. Cardiovascular complications in newly diagnosed rheumatic heart disease patients at mulago hospital, uganda. *Cardiovascular journal of Africa*, 24(3):82, 2013.

[81] A. Papolos, J. Narula, C. Bavishi, F. Chaudhry, and P. Sengupta. Us hospital use of echocardiography: insights from the nationwide inpatient sample. *Journal of the American College of Cardiology*, 67(5):502–511, 2016.

[82] A. Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

[83] F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[84] M. Ploutz et al. Handheld echocardiographic screening for rheumatic heart disease by non-experts. *Heart*, 102(1):35–39, 2016.

[85] PRB. World population data sheet 2020. Population Reference Bureau, 2020.

[86] P. Rajpurkar et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[87] A. Ran et al. Deep learning in glaucoma with optical coherence tomography: A review. *Eye*, 35(1):188–201, 2021.

[88] B. Reményi et al. World heart federation criteria for echocardiographic diagnosis of rheumatic heart diseasean evidence-based guideline. *Nature reviews cardiology*, 9(5):297, 2012.

[89] A. Ribeiro, B. Duncan, L. Brant, P. Lotufo, J. Mill, and S. Barreto. Cardiovascular health in brazil: trends and perspectives. *Circulation*, 133(4):422–433, 2016.

[90] K. Roberts, A. Brown, G. Maguire, D. Atkinson, and J. Carapetis. Utility of auscultatory screening for detecting rheumatic heart disease in high-risk children in australia's northern territory. *Medical Journal of Australia*, 199(3):196–199, 2013.

[91] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[92] R. Rosati et al. A new information system for medical practice. *Archives of Internal Medicine*, 135(8):1017–1024, 1975.

[93] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[94] G. Roth et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021, 2020.

[95] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[96] S. Russell and P. Norvig. Artificial intelligence: a modern approach. 2002.

[97] S. Sanyal, A. Berry, S. Duggal, V. Hooja, and S. Ghosh. Sequelae of the initial attack of acute rheumatic fever in children from north india. a prospective 5-year follow-up study. *Circulation*, 65(2):375–379, 1982.

[98] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[99] H. Shmueli. Briefly trained medical students can effectively identify rheumatic mitral valve injury using a hand-carried ultrasound. *Echocardiography*, 30(6):621–626, 2013.

[100] M. Shung-King, L. Zuhlke, M. Engel, and B. Mayosi. Asymptomatic rheumatic heart disease in south african schoolchildren: implications for addressing chronic health conditions through a school health service: in practice. *South African Medical Journal*, 106(8):761–762, 2016.

[101] K. Simonyan and K. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[102] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[103] A. Steer, M. Danchin, and J. Carapetis. Group a streptococcal infections in children. *Journal of paediatrics and child health*, 43(4):203–213, 2007.

[104] D. D. Sudeep and K. Sredhar. The descriptive epidemiology of acute rheumatic fever and rheumatic heart disease in low and middle-income countries. *Am J Epidemiol Infect Dis*, 1(4):34–40, 2013.

[105] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[106] S. Tavpritesh, K. Anushtha, S. Arjun, and N. Aditya. Chapter 1 - interpretable artificial intelligence: Closing the adoption gap in healthcare. In *Artificial Intelligence in Precision Health*, pages 3–29. Academic Press, 2020.

[107] D. Tompkins, B. Boxerbaum, and J. Liebman. Long-term prognosis of rheumatic fever patients receiving regular intramuscular benzathine penicillin. *Circulation*, 45(3):543–551, 1972.

[108] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. ICCV*, pages 4489–4497, 2015.

[109] S. Vakamudi et al. Gender differences in the etiology of mitral valve disease. *Journal of the American College of Cardiology*, 69(11 Supplement):1972, 2017.

[110] S. Vandenberg. Medical diagnosis by computer: Recent attempts and outlook for the future. *Behavioral Science*, 5(2):170, 1960.

[111] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, pages 1–15, 2019.

[112] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[113] H. Wang et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proc. CVPR Workshops*, pages 24–25, 2020.

[114] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks (IJCNN)*, pages 4368–4374. IEEE, 2016.

[115] E. Wark, Y. Hodder, C. Woods, and G. Maguire. Patient and health-care impact of a pilot rheumatic heart disease screening program. *Journal of paediatrics and child health*, 49(4):297–302, 2013.

[116] D. Watkins et al. Global, regional, and national burden of rheumatic heart disease, 1990–2015. *New England Journal of Medicine*, 377(8):713–722, 2017.

[117] H. Weinrauch and A. Hetherington. Computers in medicine and biology. *Journal of the American Medical Association*, 169(3):240–245, 1959.

[118] WHO et al. Task shifting: rational redistribution of tasks among health workforce teams: global recommendations and guidelines. 2007.

[119] J. Wilson and G. Jungner. Principles and practice of screening for disease. 1968.

[120] D. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[121] H. Wood, R. Simpson, A. Feinstein, A. Taranta, E. Tursky, and G. Stollerman. Rheumatic fever in children and adolescents. a long-term epidemiologic study of subsequent prophylaxis, streptococcal infections, and clinical sequelae. i. description of the investigative techniques and of the population studied. *Annals of internal medicine*, 60:6–17, 1964.

[122] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 31(2):210–227, 2009.

[123] J. Yanase and E. Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138:112821, 2019.

[124] J. Yoon et al. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. *Scientific reports*, 10(1):1–9, 2020.

[125] J. Zhang et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

[126] B. Ziaeian and G. Fonarow. Epidemiology and aetiology of heart failure. *Nature Reviews Cardiology*, 13(6):368–378, 2016.

[127] P. Zilla et al. The cape town declaration on access to cardiac surgery in the developing world. *Asian Cardiovascular and Thoracic Annals*, 26(7):535–539, 2018.

Appendix A

Information for Diagnosis Prediction

Table A.1: **Correct diagnosis, predicted diagnosis with the final model and meta-features for each exam from the sample.** The text in the *Predicted Diagnosis* column is colored to indicate a wrong (red) or correct (green) exam diagnosis prediction.

	Correct Diagnosis	Predicted Diagnosis	Confidence Mean	Confidence Std	Confidence Skewness	Confidence Kurtosis
Exam 1	RHD Negative	RHD Positive	0.658	0.266	-0.987	0.118
Exam 2	RHD Positive	RHD Negative	0.216	0.224	1.357	0.620
Exam 3	RHD Negative	RHD Negative	0.422	0.244	0.168	-1.190
Exam 4	RHD Positive	RHD Positive	0.695	0.263	-0.352	-1.117

Table A.2: **Confidence in the predicted diagnosis for RHD per video for each exam from the sample.** The first videos from each exam are the ones used in the Figure 3.7, respectively.

	# Videos	Confidence to diagnosis each video as RHD Positive
Exam 1	11	[0.973, 0.649, 0.332, 0.765, 0.046, 0.929, 0.747, 0.724, 0.900, 0.457, 0.720]
Exam 2	18	[0.031, 0.005, 0.766, 0.056, 0.014, 0.136, 0.091, 0.561, 0.127, 0.313, 0.165, 0.269, 0.260, 0.052, 0.053, 0.124, 0.176, 0.691]
Exam 3	13	[0.020, 0.735, 0.303, 0.179, 0.282, 0.460, 0.213, 0.466, 0.777, 0.176, 0.436, 0.640, 0.800]
Exam 4	19	[0.988, 0.996, 0.294, 0.622, 0.495, 0.555, 0.987, 0.664, 0.232, 0.920, 0.670, 0.997, 0.405, 0.766, 0.987, 0.994, 0.246, 0.669, 0.717]