INTEGRATING DEEP LEARNING FOR HEART AND VASCULAR ACOUSTIC ANALYSIS IN CARDIOVASCULAR HEALTH ASSESSMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

CEYDA ÖZÇİL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MECHANICAL ENGINEERING

DECEMBER 2023

Approval of the thesis:

INTEGRATING DEEP LEARNING FOR HEART AND VASCULAR ACOUSTIC ANALYSIS IN CARDIOVASCULAR HEALTH ASSESSMENT

submitted by **CEYDA ÖZÇİL** in partial fulfillment of the requirements for the degree of **Master of Science in Mechanical Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ______________

Prof. Dr. Mehmet Ali Sahir Arıkan
Head of Department, **Mechanical Engineering** ______________

Prof. Dr. Yiğit Yazıcıoğlu
Supervisor, **Mechanical Engineering, METU** ______________

Examining Committee Members:

Assoc. Prof. Dr. Ahmet Buğra Koku
Mechanical Engineering, METU ______________

Prof. Dr. Yiğit Yazıcıoğlu
Mechanical Engineering, METU ______________

Assoc. Prof. Dr. Mehmet Bülent Özer
Mechanical Engineering, METU ______________

Assoc. Prof. Dr. Hande Alemdar
Computer Engineering, METU ______________

Assist. Prof. Dr. Hüseyin Enes Salman
Mechanical Engineering, TOBB ETU ______________

Date: 07.12.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ceyda Özçil

Signature :

ABSTRACT

INTEGRATING DEEP LEARNING FOR HEART AND VASCULAR ACOUSTIC ANALYSIS IN CARDIOVASCULAR HEALTH ASSESSMENT

Özçil, Ceyda
M.S., Department of Mechanical Engineering
Supervisor: Prof. Dr. Yiğit Yazıcıoğlu

December 2023, 90 pages

Atherosclerosis, a cardiovascular disease, disrupts blood flow due to occlusions. The transformation from laminar into turbulent flow produces an acoustic phenomena known as bruits. In this study, heart sounds recorded by phonocardiography were classified as normal and abnormal using different combinations of feature extraction and classification techniques. An experiment-based model was employed to generate pulsating flow sound at different stenosis levels. Deep learning and feature comparison methodologies were applied to explore the correlation between phonocardiography and vascular sounds. Beyond promising results in heart sound classification, the study demonstrated an apparent relationship between phonocardiography recordings and 50-90% stenosed vascular sounds. This outcome highlights that coronary artery disease could be detected by utilizing the phonocardiography.

Keywords: Heart Sound Classification, Phonocardiography, Stenosis Detection

ÖZ

KARDİYOVASKÜLER SAĞLIĞIN KALP VE VASKÜLER AKUSTİK ANALİZİNE DERİN ÖĞRENME ENTEGRE EDİLEREK DEĞERLENDİRİLMESİ

Özçil, Ceyda
Yüksek Lisans, Makina Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Yiğit Yazıcıoğlu

Aralık 2023 , 90 sayfa

Kardiyovasküler bir hastalık olan ateroskleroz, tıkanıklıklar nedeniyle kan akışını bozar. Laminerden türbülanslı akışa dönüşüm, uğultu olarak bilinen bir akustik fenomene neden olur. Bu çalışmada fonokardiyografi ile kaydedilen kalp sesleri, özellik çıkarma ve sınıflandırma tekniklerinin farklı kombinasyonları kullanılarak normal ve anormal olarak sınıflandırıldı. Farklı darlık seviyelerinde pulsatil akış sesi üretmek için deney bazlı bir model kullanıldı. Fonokardiyografi ile vasküler sesler arasındaki korelasyonu araştırmak için derin öğrenme ve özellik karşılaştırma yöntemleri uygulandı. Kalp sesi sınıflandırmasında umut verici sonuçların ötesinde, fonokardiyografi kayıtları ile %50-90 darlıklı vasküler sesler arasında açık bir ilişki olduğunu ortaya koydu. Bu sonuç, koroner arter hastalığının fonokardiyografi kullanılarak tespit edilebileceğini vurgulamaktadır.

Anahtar Kelimeler: Kalp Sesi Sınıflandırılması, Fonokardiografi, Stenoz Tespiti

To my lovely family...

ACKNOWLEDGMENTS

I would like to express my endless gratitude to my supervisor Prof. Dr. Yiğit Yazıcıoğlu for his invaluable guidance, support and kindness.

I extend my thanks to the members of my committee , Assoc. Prof. Dr. Ahmet Buğra Koku, Assoc. Prof. Dr. Hande Alemdar, Assoc. Prof. Dr. Mehmet Bülent Özer and Assist. Prof. Dr. Hüseyin Enes Salman for their insightful feedback and suggestions.

I am deeply thankful to my mother, Şükriye Koruk, and my father, Adnan Koruk, for raising us with dedication, faith, and love.

My lovely hubby and my best friend İsmail Özçil, thank you for your unwavering support and delightful life companionship. Also, I sincerely thank Melahat, Celalettin and Caner Özçil for their support.

I thank my cheerleaders Sevde Sena Zeybek and Ceren Cansu Esmek for their uplifting spirits.

My special thanks go to you, my big brother Talha Koruk, for standing by me all the time and enlightening my way.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vi

ACKNOWLEDGMENTS viii

TABLE OF CONTENTS ix

LIST OF TABLES xii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvii

CHAPTERS

1 INTRODUCTION 1

1.1 Problem Definition 2

1.1.1 Atherosclerosis 2

1.1.2 Diagnosis Methods 4

1.1.3 Acoustic Phenomena in Atherosclerosis 6

1.2 Literature Survey 9

1.2.1 Analytical, Experimental and Numerical Studies 9

1.2.2 Technological Advancements 11

1.2.3 Contribution of Artificial Intelligence 13

1.3 The Purpose and the Scope of the Study 17

1.4 The Outline of the Thesis 18

2 FEATURE EXTRACTION AND CLASSIFICATION METHODS 19

2.1 Feature Extraction Methods 19

2.1.1 MFCC Heat Maps 19

2.1.2 OpenL3 20

2.1.3 VGGish 25

2.2 Classification Methods 28

2.2.1 Convolutional Neural Network 28

2.2.2 Support Vector Machine 31

2.2.3 K-Nearest Neighbor 33

3 CLASSIFICATION OF HEART SOUNDS 37

3.1 Dataset (PhysioNet Heart Sound Database) 37

3.2 Environment 41

3.3 Prediction and Performance Metrics Definition 42

3.4 Algorithm Implementations 42

3.4.1 MFCC Heat Maps 42

3.4.2 OpenL3 44

3.4.3 VGGish 46

3.4.4 Convolutional Neural Network 47

3.4.5 Support Vector Machine 49

3.4.6 K-Nearest Neighbor 51

3.5 Experiments 52

3.5.1 MFCC and CNN 52

3.5.2 OpenL3 and SVM 56

3.5.3 OpenL3 and KNN 58

3.5.4 VGGish and SVM 60

3.5.5 VGGish and KNN 62

3.6 Discussion of the Results for Chapter 3 64

4 EXPLORING THE RELATION BETWEEN PHONOCARDIOGRAPHY AND VASCULAR SOUNDS 73

4.1 Vascular Sound Dataset 73

4.2 Relating Heart and Vascular Sounds 79

4.2.1 Evaluating with Pre-trained MFCC and CNN 79

4.2.2 Evaluating with Feature Comparison 80

4.3 Discussion of the Results for Chapter 4 81

5 CONCLUSION AND FUTURE WORK 83

5.1 Conclusion 83

5.2 Future Work 84

REFERENCES 85

LIST OF TABLES

TABLES

Table 2.1 Hyperparameter Values 31

Table 2.2 Network Parameter Values 31

Table 3.1 Distribution of Normal and Abnormal Sizes Across Different Categories 41

Table 3.2 The Number of Record and Feature Attempts Summary 53

Table 3.3 Learning Rate vs. Validation Accuracy Comparison 53

Table 3.4 New Hyperparameter Values 54

Table 3.5 New Network Parameter Values 54

Table 3.6 Validation Results Before the Training 55

Table 3.7 Validation Results After the Training 55

Table 3.8 Test Results After the Linear Kernel SVM Classification 57

Table 3.9 Test Results After the Polynomial Kernel SVM Classification 57

Table 3.10 Test Results After the RBF Kernel SVM Classification 58

Table 3.11 Test Results After the Euclidian Distance KNN Classification 59

Table 3.12 Test Results After the Manhattan Distance KNN Classification . . . 59

Table 3.13 Test Results After the Minkowski Distance KNN Classification . . . 60

Table 3.14 Test Results After the Linear Kernel SVM Classification 61

Table 3.15 Test Results After the Polynomial Kernel SVM Classification 61

Table 3.16 Test Results After the RBF Kernel SVM Classification 62

Table 3.17 Test Results After the Euclidian Distance KNN Classification 62

Table 3.18 Test Results After the Manhattan Distance KNN Classification . . . 63

Table 3.19 Test Results After the Minkowski Distance KNN Classification . . . 63

Table 3.20 Summary of OpenL3 and SVM Results with Different Kernels . . . 64

Table 3.21 Summary of OpenL3 and KNN Results with Different Distance Metrics 66

Table 3.22 Summary of VGGish and SVM Results with Different Kernels . . . 68

Table 3.23 Summary of VGGish and KNN Results with Different Distance Metrics 69

Table 3.24 Summary of the Best Results Across Different Methods 70

Table 4.1 List of Pressure Distribution Equation's Parameters 74

Table 4.2 Summary of Pressure Distribution Generation Parameters 75

Table 4.3 Summary of Sound Generation Parameters 80

LIST OF FIGURES

FIGURES

Figure 1.1 Evolution of Atherosclerotic Pathogenesis 2

Figure 1.2 Stages in the Development of Atherosclerotic Lesions 3

Figure 1.3 Stenosis Flow Diagram 8

Figure 1.4 Acoustic Vibration Generation of Atherosclerotic Artery 8

Figure 1.5 Cross-sections of Occluded Coronary Arteries in Different Morphology 10

Figure 1.6 The Prototype of Multimodal Patch 13

Figure 1.7 Overview of the Proposed System 14

Figure 1.8 Final Scores of the Best 8 Entrants 15

Figure 1.9 The Network Architecture of the HeartFit 16

Figure 1.10 The Workflow of ML Audio Classification 16

Figure 2.1 Visualization of MFCC Calculation Process 19

Figure 2.2 Illustration of Audio-Visual Correspondence Method 20

Figure 2.3 L^3-Net Arhitecture 21

Figure 2.4 Classification Accuracy Against Input Representation 22

Figure 2.5 Classification Accuracy Against Different Training Datasets . . . 23

Figure 2.6 Comparison of Classification Accuracy Relative to the Number of Training Samples in Embedding Model Training 25

Figure 2.7 VGGish Network Architecture 27

Figure 2.8 Analogy between the Human and Artificial Neuron 28

Figure 2.9 Simple Neural Network Representation 29

Figure 2.10 Convolutional Neural Network 30

Figure 2.11 Visual Representation of SVM Basic Concepts 32

Figure 2.12 Illustration of KNN Classification Approach 35

Figure 3.1 Optimal Auscultation Points 40

Figure 3.2 Illustration of Sample Rate Decrease by Segmentation Algorithm 43

Figure 3.3 Illustration of MFCC Heat Map Process Summary 43

Figure 3.4 Normal(Healthy) Audio Sample Waveform 44

Figure 3.5 Normal Sample MFCC Heat Map 44

Figure 3.6 Abnormal Audio Sample Waveform 44

Figure 3.7 Abnormal Sample MFCC Heat Map 44

Figure 3.8 Illustration of OpenL3 Algorithm Implementation 45

Figure 3.9 Illustration of VGGish Algorithm Implementation 46

Figure 3.10 Illustration of CNN Algorithm Implementation 48

Figure 3.11 Illustration of SVM Implementation 49

Figure 3.12 Illustration of KNN Implementation 51

Figure 3.13 The New Convolutional Network Architecture 54

Figure 3.14 Epoch Number vs. Accuracy Graph 56

Figure 3.15 ROC Curve for OpenL3 and SVM with Different Kernels 65

Figure 3.16 ROC Curve for OpenL3 and KNN with Different Distance Metrics 67

Figure 3.17 ROC Curve for VGGish and SVM with Different Kernels 68

Figure 3.18 ROC Curve for VGGish and KNN with Different Distance Metrics 70

Figure 3.19 ROC Curve for the Best Results Across Different Methods . . . 71

Figure 4.1 RMS Wall Pressure Fluctuations wrt x/D 73

Figure 4.2 Comparison between the Different Stenosis Severities 75

Figure 4.3 Comparative Visualization of Acoustic Characteristics for a 70% Stenosis Condition 78

Figure 4.4 Comparative Visualization of Acoustic Characteristics for a 90% Stenosis Condition 78

Figure 4.5 Schematic illustration of the evaluation of vascular sounds using a pre-trained MFCC and CNN 79

Figure 4.6 Schematic Illustration of the Cross-Correlation Process Using OpenL3 Features of Phonocardiography and Vascular Sounds 81

Figure 4.7 The Stenosis Level vs. Abnormal Probability Plot 82

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AVC	Audio-Visual Correspondence
CAD	Coronary Artery Disease
CNN	Convolutional Neural Network
DL	Deep Learning
ECG	Electrocardiogram
FP	False Positives
FN	False Negatives
KNN	K-Nearest Neighbors
MFCC	Mel-frequency Cepstral Coefficients
ML	Machine Learning
PCG	Phonocardiogram
Re	Reynolds Number
ROC	Receiver Operating Characteristic
Se	Sensitivity
Sp	Specificity
SVM	Support Vector Machines
TP	True Positives
TN	True Negatives

CHAPTER 1

INTRODUCTION

Non-communicable diseases (NCDs) are commonly called chronic diseases, including cancers, diabetes, respiratory and cardiovascular diseases which have long-term consequences. They arise from lifestyle, genetic factors, and environmental conditions. Excessive alcohol intake, tobacco consumption, physical inactivity, and bad eating habits increase the probability of developing diseases. According to World Health Organization (WHO), the prominent mortality rate percentage belongs to cardiovascular diseases, and they mention that the number of deaths is approximately 17.9 million annually [1].

Atherosclerosis, a type of cardiovascular disease, is a significant contributor to these deaths. It is a condition that progressively causes the buildup of substances in artery walls resulting in plaque formation. These plaques can eventually limit blood flow to essential organs such as the heart, brain, and kidneys. It leads to vital health issues like heart attacks and strokes. Contrary to what was known before, the prevalence of atherosclerotic cardiovascular diseases is no longer confined to industrialized nations but it has escalated into a worldwide concern [2].

The following section will provide a detailed description, blockage mechanism, and diagnosis methods of atherosclerosis. It is essential to clearly understand these aspects to develop effective approaches to detect and treat this common and potentially lethal disease. After clarifying the disease, there will be a literature survey section to gain insight from the latest studies.

1.1 Problem Definition

1.1.1 Atherosclerosis

Atherosclerosis has a Greek-based etymology, which means hardening of the vessels. Although cholesterol is the leading cause of atherosclerosis, there is a more complicated mechanism than commonly known.

The initial growth of atherosclerotic lesions are referred to as atherogenesis or pathogenesis. Figure1.1 shows the evolution of pathogenesis progress in years. It was considered a lipid storage-based disease in the beginning. Through further research and investigation, it is realized that immune response, endothelial dysfunction, and inflammation could be the reason for the plaque accumulation [3].

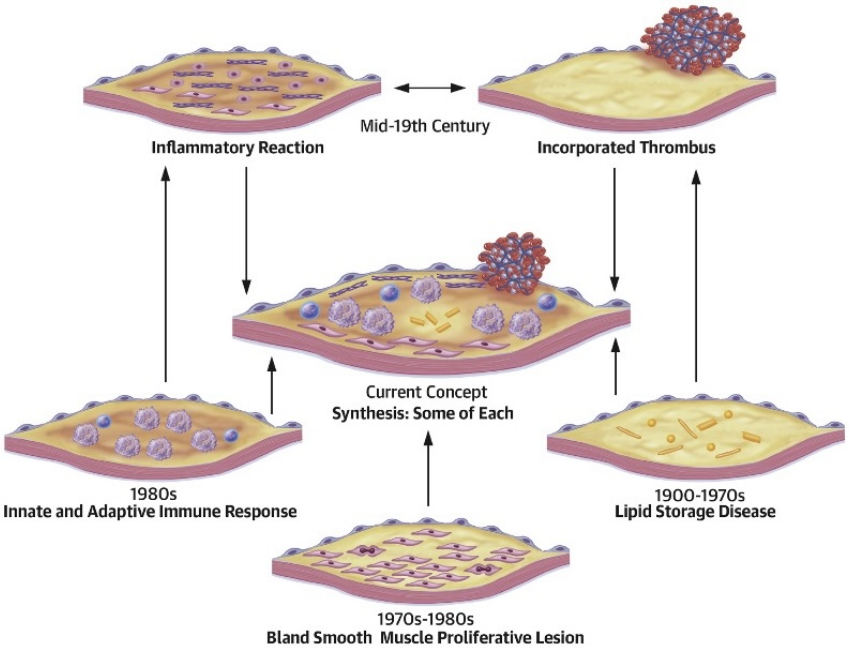


Figure 1.1: Evolution of Atherosclerotic Pathogenesis adopted from [3]

Figure 1.2 is an illustration of the stages in the development of atherosclerotic lesions. The disease starts with damage to the inner layer of an artery called the endothelium. Normally, this layer prevents molecules to pass through to the inner layers in healthy conditions. However, high blood pressure, high cholesterol levels, or inflammation can weaken the integrity of the endothelium layer. This makes it more permeable, allowing lipids and immune system molecules to enter the artery wall.

Low-density lipoprotein (LDL), known as bad cholesterol, penetrates through the damaged surface and accumulates in the artery wall's intima layer. Also, the monocytes in the blood attach to the endothelium, pass through the intima layer and turn into macrophages. The new form of monocytes swallows the LDL molecules. They evolve to foam cells and create fatty streaks after accumulation.

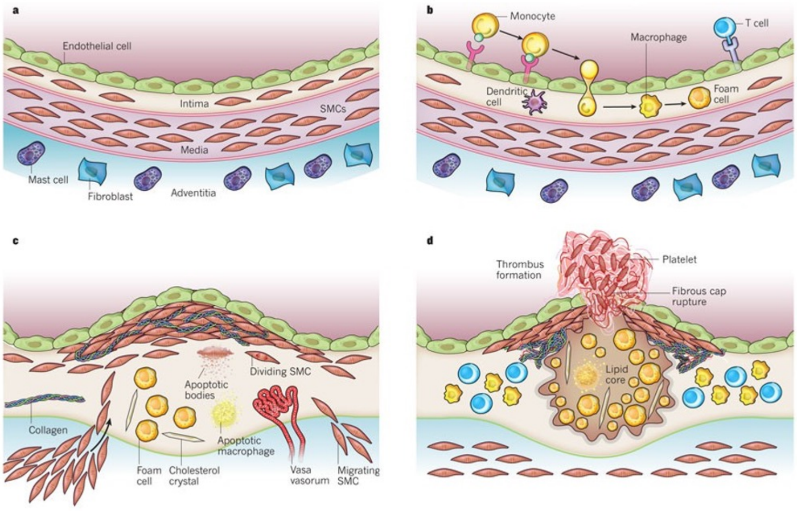


Figure 1.2: Stages in the Development of Atherosclerotic Lesions adopted from [4]

The long-term development of the streaks results in atheromatous plaques with a lipid-rich core and a fibrous cap. Depending on unstable fibrous surfaces, vascular plaques are more prone to rupture. When a plaque bursts, a lipid-rich content mixes into the bloodstream, which triggers thrombosis(blood clots) formation.

Inflammation plays a critical part in the entire process. Plaques cause an immunological reaction, which attracts inflammatory cells, such as T cells, to the artery wall and aids in the growth and instability of plaques.

As a result, the blood flow regime is altered by expanding plaques and narrowing the artery lumen. Accordingly, the vital organs cannot receive enough oxygen from the blood. If vessels feed the heart, these conditions lead to angina or myocardial infarctions. In carotid arteries, it can increase the probability of strokes.

This disease progresses gradually at a slow pace. Typically, atherosclerosis does not produce symptoms until an artery is either completely blocked or significantly narrowed [5]. There can be irreversible effects on the vital functions of humans, or it may result in death. Therefore, diagnosing and treating atherosclerosis immediately will have remarkable results.

1.1.2 Diagnosis Methods

The early diagnosis of atherosclerosis has the utmost importance in preventing irreversible and permanent consequences. The methods focus on assessing narrowing or occlusion in the artery lumen. There are standard methods to diagnose the disease, such as invasive and noninvasive techniques. The decision depends on where the blockage is suspected, the patient's health status, and the severity of the conditions.

The most reliable and suitable method for high-risk patients is coronary angiography which is in vivo [6]. The procedure employs a special dye and X-rays. The dye, visible by X-ray, is injected into the blood from the groin or arm of the patient with a catheter, so it is an invasive method. The image retrieved from the X-ray shows the details about narrowing, occlusion, or abnormalities of arteries. The technique has excellent spatial and temporal resolution but could not contribute information about small lesions [7].

The next gold standard techniques seem like Intravascular Ultrasound (IVUS) and Optical Coherence Tomography (OCT). They both have better performance to qualify the artery dimensions [8]. IVUS employs sound waves to evaluate soft vascular tissue. A tiny piece of equipment captures real-time images piece by piece from veins and is mainly used for coronary arteries. Similarly, OCT is a tool to visualize lesions of veins with microscopic precision by the light beams to evaluate the tissue rather than sound [9]. If there is a lipid accumulation with high cholesterol, IVUS, and OCT are not good at imaging plaques.

Near Infrared Spectroscopy (NIS) is an alternative intravascular imaging method in this case. Since cholesterol molecules absorb infrared light, occlusion appears from the yellow to red spectrum. The practitioners can combine these intravascular methods to utilize superior properties for accurate and precise results [10].

Intravascular methods are not only helpful in imaging the artery and discretization of different lesions, but also it is useful for applications like stent implantation. Although these methods are quite reliable, they have limitations and risks to be considered [11].

Firstly, Atherosclerotic diseases should be diagnosed to treat as soon as possible, but intravascular techniques have cumbersome procedures that take a long time. The next disadvantage is that qualified professionals must handle the operations. It is also necessary to consider the initial costs of imaging devices that affect patients and medical staff by emitting radiation. In addition to them, there can occur clinical complications. Inserting a catheter into the blood vessel damages the vessel wall. According to coronary health, the effect of damage can be either minor or significant. Paradoxically, these techniques could cause the disease trying to detect by triggering the formation of blood clots or leading to the spread of plaques through the body and leading to another occlusion. Further, blood pressure drops could change normal heart functioning. Also, bruises on the skin surface have always been a risk source of infections. Therefore, taking preventive cautions to avoid risks and evaluating conditions thoroughly before application is essential.

The risks of invasive methods and the silent progression of the disease call for new noninvasive instruments. Magnetic Resonance Imaging (MRI) Angiography has a different procedure than conventional invasive angiography. Although a contrast dye is injected into the vessel, there is no catheterization. It employs powerful magnets and radio waves to create high-resolution blood vessel images.

Computed Tomography Angiography (CTA) can also image the cross-sectional areas of blood vessels with various images collected from distinct body angles. This technique is more applicable in detecting calcium accumulation and provides more information than direct X-ray images due to computer support [12].

Conventional ultrasound applications are very rapid and cost-efficient solutions. Doppler ultrasound is a type that uses sound waves to distinguish the speed and direction of blood flow, which identifies areas of narrowing, blockages, and the presence of atherosclerotic plaques. Bright mode ultrasound can visualize the thickness and shape of the artery as well as the presence of abnormalities. Combining these two methods results in a comprehensive image of the vascular system [13].

Non-invasive methods also carry risks and potential dangers, just like invasive ones. Although they are relatively safer, being aware of any potential risks is crucial. The most prominent consideration is the radiation exposure from X-ray-based tool because it has an ionizing nature that can disrupt normal chemical reactions and damage the cellular structure. Prolonged or frequent exposure leads to cell mutations, tissue deterioration, and, eventually, cancers. The other aspect is the contrast agents. Despite the image-enhancing effect of dyes, they can harm kidneys and exacerbate preexisting conditions. Also, an allergic reaction can occur, even if it is rare. In addition to the high initial cost and the necessity for skilled professionals, non-invasive methods result in a false-negative or false-positive case.

The last but not the least non-invasive diagnosis method is phonocardiography. The technique involves recording and analyzing heart sounds. It is able to identify abnormalities that may indicate underlying cardiovascular issues. Phonocardiography offers a non-invasive, clear, and cost-effective means to evaluate cardiac functions by catching distinct sound patterns. Since the process is very proper for data collection, machine learning algorithms can utilize them to detect abnormalities without well-skilled professionals. This approach complements other diagnostic tools and helps clinicians in early detection, accurate assessment, and punctual intervention for coronary artery disease. In this way, this method reduces potential complications and enhances the healthcare management.

1.1.3 Acoustic Phenomena in Atherosclerosis

Under normal conditions, blood flow is laminar. It moves smoothly and steadily. However, when blood encounters a narrowing in a blood vessel, this laminar flow is disrupted, which causes a significant pressure drop within the artery. This pressure drop is particularly critical in the coronary artery system, where there is a direct proportional relationship between the drop in pressure and the heart's energy demands. A higher pressure drop means the heart's muscle must work harder to maintain a consistent blood flow to the organs. The pressure decrease through a simplified representation of the occluded coronary artery model can be expressed with following expression:

$$\Delta P_{\text{blockage}} = f\left(Re, \frac{L}{D}, \frac{A_1}{A_0}, \frac{l}{D}, e, \frac{dV}{dt}\right) \tag{1.1}$$

In the equation labeled (1.1), ΔP stands for the pressure drop in an artery with a specific length L and internal diameter D. The term Re refers to the Reynolds number, expressed by equation 1.2, is a measure of the flow type of the blood. The areas A_0 and A_1 are the cross-sectional areas of the artery where there is no blockage and the narrowest part of the blockage, respectively. The length of the blockage is given by l, and e is its eccentricity, which describes how off-center the blockage is. Lastly, V is the speed of the blood flow in the artery without blockage.

$$Re = \frac{\rho V_{avg} D}{\mu} \tag{1.2}$$

Reynolds number(Re) involves the density of the blood (ρ), the average velocity (V_{avg}) in the part of the artery without any blockage, and the blood's viscosity (μ).

The blood flow Re range in the body varies from 1 in small arterioles to approximately 4000 in the largest artery. At a Reynolds number of 2000, steady laminar flow in a circular pipe becomes unsteady. Transition into fully turbulent flow occurs at approximately 4000. [14] Nevertheless, Yongchareon and Young [15] proposed that turbulence could emerge at even a lower Reynolds number due to flow disruptions.

Fredberg [16] proposed that, right after the occlusion site, the blood flow separates from the walls since it can not overcome the adverse pressure gradient. The consequence of this detachment is the creation of a high-velocity jet stream. A shear layer is created when this jet stream interacts with the slower recirculating fluid in the recirculating separation zone. This particular layer is highly susceptible to fluid-dynamic instabilities, which are commonly referred to as shear instabilities. Within this shear layer, turbulence occurs as the instabilities grow and extract energy from the mean flow. The visual representation of the blood flow through the obstructed region is given in Figure 1.3.

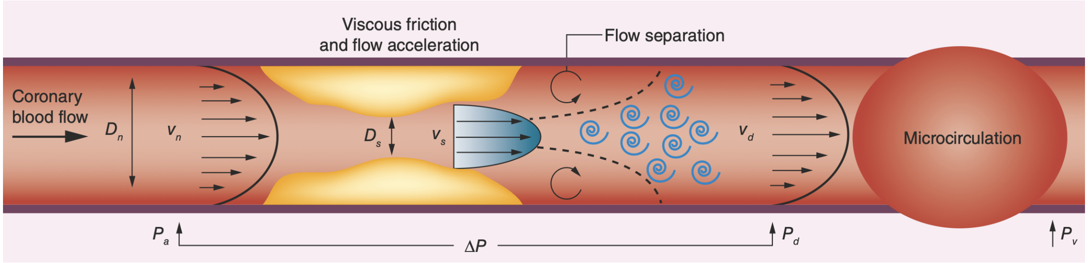


Figure 1.3: Stenosis Flow Diagram retrieved from [17]

The resulting pressure fluctuations cause tissue vibrations and generate vascular sounds. Figure 1.4 shows how turbulence after a stenosis generates sound and how this sound is transmitted to the skin surface. These sounds were first attributed as bruits by René Laennec, who invented the stethoscope. Lees and Dewey [18] introduced a new diagnostic approach depending on these acoustic vibrations, *Phonoangiography*. They suggested that these waves on the body's surface can be detected using tools such as a stethoscope or a device that measures skin movement. By doing this, we can measure and record these sounds' strength and range of frequencies. The use of quantitative analysis can help investigate fluid movement in atherosclerotic arteries.

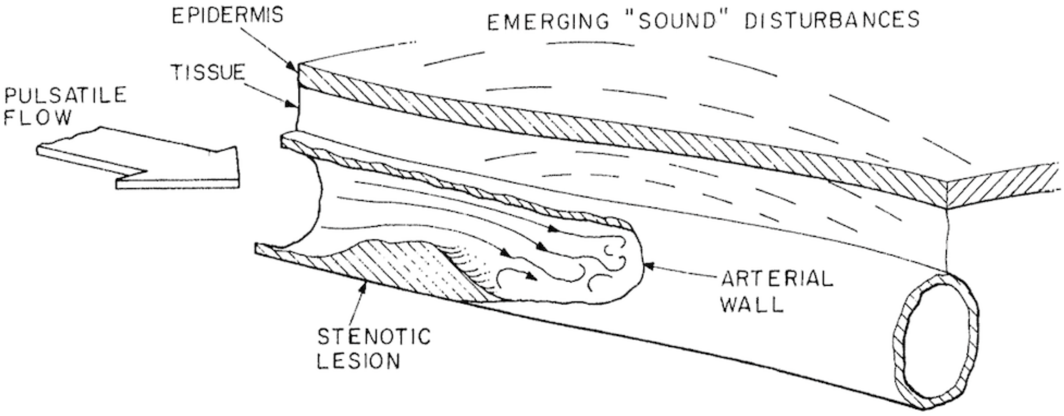


Figure 1.4: Acoustic Vibration Generation of Atherosclerotic Artery retrieved from [18]

1.2 Literature Survey

The practice of listening to the heart, known as auscultation, dates back to the ancient civilizations of Greece and Egypt. Laennec invented the stethoscope in 1816, and it was the initialization of dedicated research into understanding heart sounds. [19] There have been many fundamental studies that explain the theoretical background of our research, as well as the technological advancements that have encouraged further study in this field. The valuable studies can be categorized as analytical, experimental, and numerical studies, technological advancements, and the contributions of artificial intelligence.

1.2.1 Analytical, Experimental and Numerical Studies

The multifaceted approach to understanding the dynamics of coronary artery diseases contains analytical, experimental, and numerical studies. Each of them gives unique insights into the complexities of blood flow and associated pathologies.

Yazıcıoğlu et al. [20] analyze vibrations in a thin-walled, viscoelastic tube experiencing turbulent flow caused by an axisymmetric constriction. They aim to provide insights into the dynamics of vascular systems and enhance noninvasive diagnostic techniques through acoustic measurements. The experiment setup mimics the conditions of a vascular system with internal fluid dynamics influenced by a constriction. It resembles an occlusion in a blood vessel. In addition to the experiment, an analytical and theoretical model was studied to analyze how turbulence in the tube causes vibrations in the surrounding materials. These results were verified against experiments using Laser Doppler Vibrometry (LDV), which measures how much the tube and the surrounding materials vibrate. Moreover, the pressure of the fluid inside the tube is also measured. Although the experimental results generally aligned with the theoretical model, some discrepancies originated from the linear structural models and the nonlinear characteristics of tissues and turbulence.

Tobin and Chang [21] have an experimental study that measures the pressure on the walls of a tube at different points after placing cylindrical blocks inside it, which mimics the blood flow in stenosed vessels. Their goal was to identify the severity of blockages in blood vessels by analyzing the sounds these blockages make, so they utilized different sizes of blockages with a constant flow of water. Particularly at high flow rates, they noticed a jet of water shooting from the narrowest part of the blockage and merging back to the tube wall within a short distance. They discovered consistent patterns between the frequency and intensity of the pressure changes and the blockage size. By introducing new variables, they standardized the pressure fluctuation data at the point of highest difference. They also compared this standardized data with the root mean square (RMS) pressure values and noticed that they were similar but not exactly the same.

Plaque morphology could be described by the artery shape and degree of blockage. Freidoonimehr et al. [22] suggested that the morphology considerably influence flow behavior and pressure drop. The stenosis degree is defined as the percentage of the arterial cross-section blocked due to plaque formation. The plaque can occupy either the entire artery cross section in a severe case or just a tiny part. Stenosis can be characterized by its geometrical shape, eccentricity, and edge sharpness. Figure 1.5 shows circular stenosis from a to e, elliptic stenosis from f to j and sharp-edge stenosis from k to o.

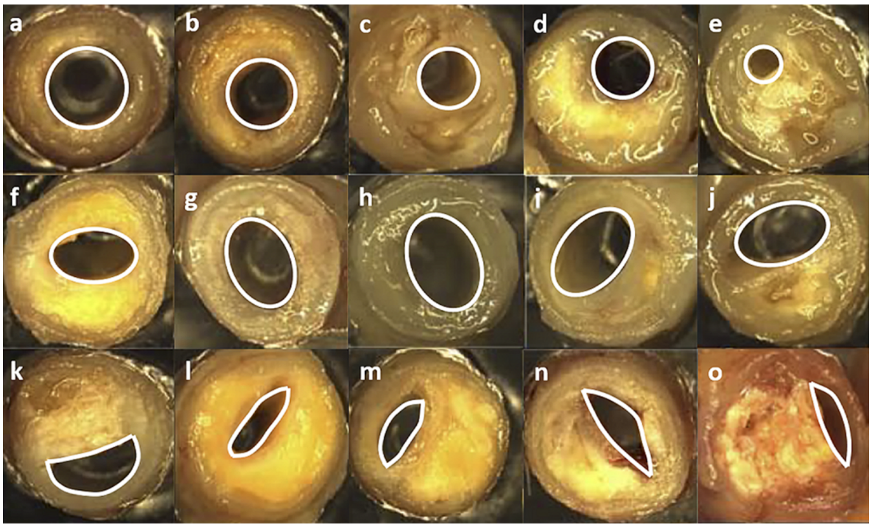


Figure 1.5: Cross-sections of Occluded Coronary Arteries in Different Morphology retrieved from [22]

Salman [23] conducted computational analyses and experimental studies to find the relationship between stenosis level, vessel parameters and corresponding vibration responses. Arteries, blood, muscles, fat, and bones have been modeled in computational study. The study investigates the effects of turbulence-induced dynamic pressure fluctuations on the arterial wall by means of the radial displacement, velocity, and acceleration responses on the skin surface. To perform this task, different flow rates, stenosis severities, and structural material properties are considered. The results obtained from the computational analysis align well with the experimental observations.

Vibrations on phantom tissue were measured using a microphone, electronic stethoscope, and laser Doppler vibrometer. As a result, a 70% blockage level is a significant threshold, as levels above 70% have shown a marked increase in vibration amplitudes. Moreover, if the blockage level increases from 70% to 90%, then the vibration amplitudes on the outer surface of the artery increase by more than tenfold.

In their numerical analysis, Ozden et al. [24] explored how the shape of a stenosis affects pressure oscillations and sound emissions in stenosed blood vessels. OpenFOAM is used to perform Large Eddy Simulations (LES) under pulsatile flow conditions with a non-Newtonian fluid blood model. The results show that a sharp rise at the start of the blockage and overlapping blockages can make the blood flow more chaotic right after the blockage, creating more swirls and energy in the flow and making the pressure change more dramatically. These blockage characteristics also make the sound louder, especially during the heart's systolic phase. However, if the blockage is uneven, it has the opposite effect. The pressure on the blood vessel walls also shows that the blockage's shape changes how loud and what pattern the heart murmurs have. This study demonstrates that the shape of the blockage is a crucial detail in generated sound.

1.2.2 Technological Advancements

There has been a significant evolution in technological capabilities in the pursuit of non-invasive methodologies for monitoring heart health. Traditional auscultation methods provided the foundation for understanding cardiac acoustics. Building upon this, advancements in acoustic technology and analyzing methods have led to the development of sophisticated devices capable of capturing cardiac sounds with precision beyond the human ear. These devices leverage digital signal processing to analyze heart sounds. Also, recent innovations have led to the miniaturization and optimization of acoustic sensors, allowing their integration into wearable devices. These wearables offer continuous monitoring of heart sounds in real-world settings. The evolution of this technology reflects a collaborative effort across decades, with incremental improvements contributing to the current state of the art.

Heart sounds originate from specific thorax locations and travel at a unique speed through different body tissues. Cardiac acoustic mapping leverages the spatial details captured on the chest surface to understand heart sound origins and pathways. Stethovest, designed by Sapsanis et al. [25], is a wearable device tailored for the upper body. The vest is equipped with 12 PCG sensors (microphone array), enabling it to capture heart sounds from multiple locations simultaneously and map the heart's acoustic activity. It is especially useful for doctors to detect abnormalities or issues with the heart by comparing sounds from different parts of the thorax all at once.

Klum et al. [26] presented a pioneering device for auscultation. It is a wearable patch that uses Bluetooth 5.0 LE to combine five different functions: a micro electromechanical system (MEMS) stethoscope, noise detection, ECG, impedance pneumography (IP), and 9-axial actigraphy. Its key benefit is its ability to replace multiple separate sensors, making long-term health monitoring more accessible and comfortable. This patch is especially useful in monitoring patients after surgery and during sleep studies. It helps identify important health events, enhances patient comfort, and reduces costs. Typically, monitoring requires several sensors placed all over the body, and the procedure can be inconvenient and uncomfortable. This compact patch (70 mm by 60 mm) integrates all these functions into one small, easy-to-wear device. Figure 1.6 demonstrates the prototype of the device. The patch has proven to be highly effective in monitoring heart and lung sounds and recording ECG and IP signals. It accurately measures breathing patterns and correlates closely with standard references. This indicates its precision and suitability for situations where discreet yet high-quality monitoring is needed. Its ability to process data at high speeds and synchronize it online further improves its function and reliability.

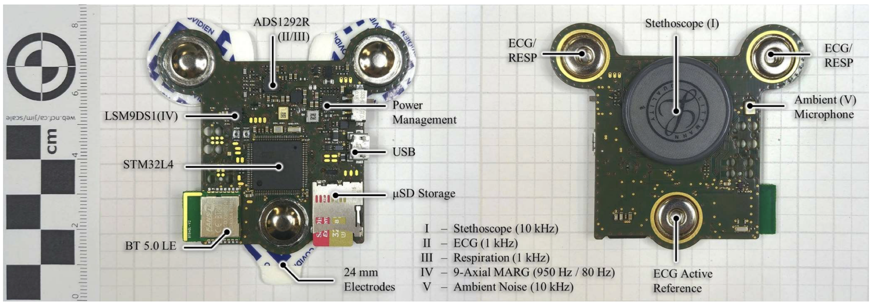


Figure 1.6: The Prototype of Multimodal Patch retrieved from [26]

Sensor technologies have become increasingly significant for accurate and precise measurements. Illustrating this, Jiangong et al. [27] introduced a unique MEMS auscultation sensor inspired by the human ear's auditory system. This sensor mimics how sound waves travel to the basement membrane through the eardrum and are transformed into nerve impulses by hair cells with cilia bundles. It features a cantilever beam with an embedded piezoresistor, simulating the cilia's role. When heart sound vibrations are detected, the beam deforms, and the piezoresistors convert this deformation into a differential voltage signal, mirroring the auditory system's function auditory system.

1.2.3 Contribution of Artificial Intelligence

Artificial intelligence (AI) is a powerful catalyst. It significantly accelerates the processing and interpretation of complex data across various fields. In cardiology, non-invasive acoustic measurement techniques integrated with AI transform CAD detection and monitoring. AI can uncover subtle patterns in heart sounds that may indicate underlying conditions, enhancing diagnostic accuracy and leading to more timely and effective treatments by harnessing sophisticated algorithms. AI and acoustic diagnostics stand on the verge of a new era in cardiology. It will complement clinical expertise and increase the successful results.

Machine learning is an assertive branch of artificial intelligence. Promising studies have been conducted using machine learning-based techniques in recent years. Samanta et al. proposed a multi-channel PCG-based system that simultaneously acquires four different auscultation areas on the thorax [28]. ANN is utilized to classify signals based on five features extracted from time and frequency domains. The process could be divided into four subsequent parts. It starts with data acquisition from 66 male subjects. These subjects were separated into two groups: healthy and CAD-positive, confirmed by angiogram. A low-pass filter and heart rate computation are applied for pre-processing the dataset. After examining the dataset in both time and frequency domains, classification is performed using selected features. The overview of the procedures is given in Figure 1.7. The proposed method achieved an accuracy of 82.57%, compared to 68.93% for the baseline CAD detection system using single-channel data. The performance enhanced considerably with the use of the multi-channel framework. The importance of this work is an affordable and safe diagnostic method that provides an augmented screening of patients with high-risk conditions.

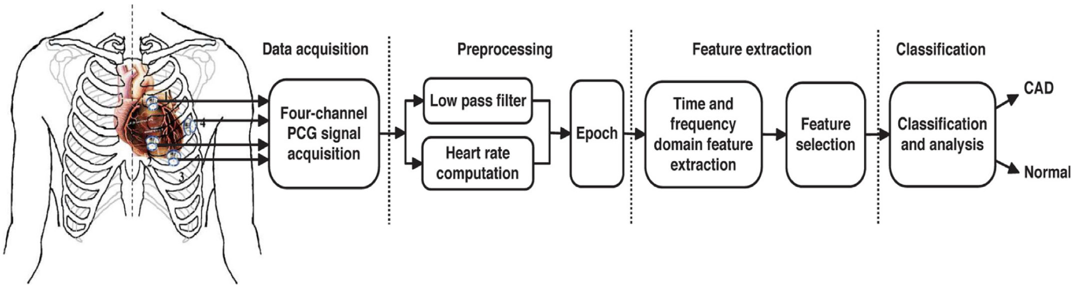
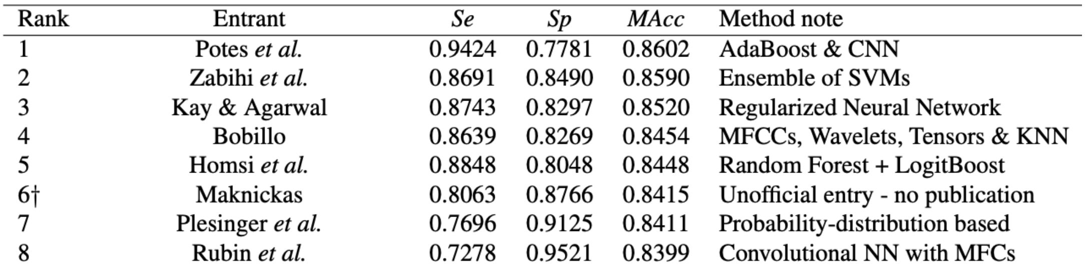


Figure 1.7: Overview of the Proposed System retrieved from [28]

The evolution of machine learning algorithms in the diagnosis of CAD has been extensively reviewed for the period spanning from 1992 to 2019 by Alizadehsani et al. [29] This review elucidates that machine learning integrated CAD diagnosis varies regarding the data types, sampling, feature selection, data collection location, performance measurement technique, and algorithm. Moreover, it emphasizes that deep learning algorithms will have a revolutionary effect on CAD detection. As deep learning models require massive datasets, there was not worthwhile research until 2016 because there is insufficient heart-sound data for deep learning.

2016 is the year that PhysioNet arranged a computing cardiology challenge [30]. This challenge is the milestone of deep learning applications in heart sound classification. PhysioNet provided 90,000 samples for training and validation. In addition to the massive dataset, they contribute a pre-processing algorithm for filtering and segmenting heart sound signals. Forty-eight teams brought unique methodologies and enriched the literature with distinct perspectives and approaches. Figure 1.8 contains 8 top entrants, scoring and their methods.



Rank	Entrant	*Se*	*Sp*	*MAcc*	Method note
1	Potes *et al.*	0.9424	0.7781	0.8602	AdaBoost & CNN
2	Zabihi *et al.*	0.8691	0.8490	0.8590	Ensemble of SVMs
3	Kay & Agarwal	0.8743	0.8297	0.8520	Regularized Neural Network
4	Bobillo	0.8639	0.8269	0.8454	MFCCs, Wavelets, Tensors & KNN
5	Homsi *et al.*	0.8848	0.8048	0.8448	Random Forest + LogitBoost
6†	Maknickas	0.8063	0.8766	0.8415	Unofficial entry - no publication
7	Plesinger *et al.*	0.7696	0.9125	0.8411	Probability-distribution based
8	Rubin *et al.*	0.7278	0.9521	0.8399	Convolutional NN with MFCs

Figure 1.8: Final Scores of the Best 8 Entrants retrieved from [30]

Rubin et al.[31] utilized a simple yet effective approach to the challenge. They combine the mel-frequency cepstral coefficients(MFCC) method for feature extraction and the convolutional network for classification. This approach inspired the thesis course and helped to gain insight into applying deep learning algorithms. After the competition, the number of studies increased based on the provided datasets.

Earlier strategies to detect abnormalities depend on support vector machines(SVM), ANN, and signal processing in various ways. These techniques have low performance, around the 80s. Gupta et al. introduced The HeartFit, a novel algorithmic approach and an innovative platform [32]. The network architecture,shown in Figure 1.9, is built with seven convolutional and five recurrent layers to classify heart sounds. It has better performance and more balanced results than the PhysioNet challengers. This success is because RNN improves performance because the sound is a sequential data type. The platform has three components: a mobile application, a database, and a deep learning server. The process involves capturing audio through a stethoscope interfaced with a mobile application. Then, it synchronizes with a deep learning server to process and exchange audio data and diagnostic results. HeartFit aims to deliver murmur monitoring available for people without access to medical devices and sufficient professional care. Moreover, it can be the option of a home murmur-monitoring system. It allows people to identify a murmur with no experience.

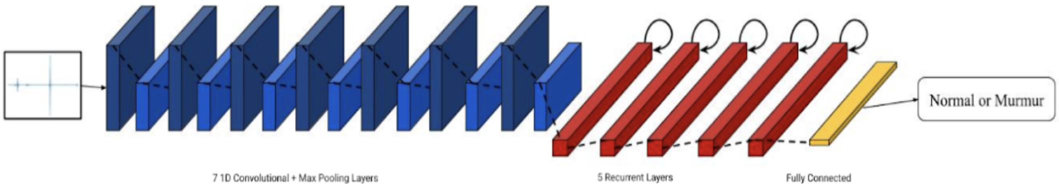


Figure 1.9: The Network Architecture of the HeartFit retrieved from [32]

Sharma et al. [33] illustrated the basic steps of sound classification with ML with Figure 1.10. The features determine the model performance of a ML algorithm. Hence, extracting features is an important component of the workflow.

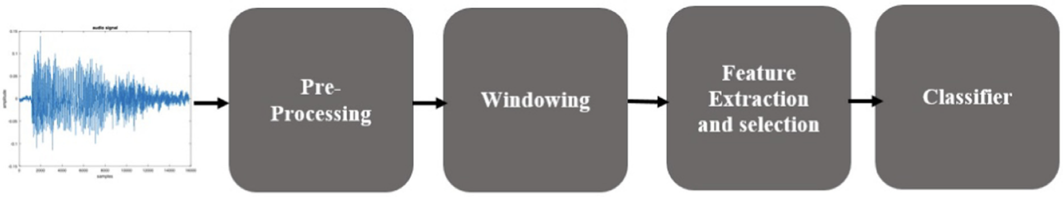


Figure 1.10: The Workflow of ML Audio Classification retrieved from [33]

Lately, there have been many studies that train audio classification models on big datasets [34, 35]. This training helps them develop complex features known as embeddings. They are useful for classification without additional domain knowledge, even for smaller datasets.

Despotovic et al. [36] utilized different feature extraction methods and compared them in their COVID-19 study. They also introduce a dataset including cough and breathing recordings from patients and healthy people. This disease can cause unique vocal patterns due to changes in breathing and voice. They employed standard audio features, VGGish, and OpenL3 to extract features and get the patterns. Although the combination of MLP and Wavelet reached 88.52% accuracy, the OpenL3 and VGGish also had promising results, approximately 76%.

The other aspect of this study is the comparison of the different supervised classification methods of the extracted features. By these methods, the algorithms allow us to learn the relationships between the features of labeled inputs and the outputs.

The primary goal is to predict the unseen instances with the model learned from the training data.

Shetha et al. [37] noted that several studies have been conducted to evaluate different algorithms for finding the best approach. These have revealed that no single solution is effective in every case. They also performed a comparative study using different datasets and focused on the four well-known classifiers: decision trees, KNNs, SVMs, and Naive Bayes. Each model achieved distinct results according to the dataset characteristics. Therefore, applying different methods to achieve the best results in specific topics is beneficial.

In conclusion, these studies demonstrate the evolution of cardiac auscultation from basic stethoscopes to AI-integrated diagnostic tools. Research ranging from analytical, experimental, and numerical studies to technological advancements and AI contributions has collectively enhanced our ability to detect and analyze cardiovascular diseases more accurately and non-invasively. The diversity of methodologies and technologies underscores the complexity of cardiac diagnostics and highlights the potential for further innovations.

1.3 The Purpose and the Scope of the Study

Atherosclerosis presents significant diagnostic challenges due to the cumbersome, costly, and invasive nature of conventional methods. This study aims to pioneer a non-invasive approach for detecting coronary artery diseases using phonocardiography recordings and artificial intelligence, utilizing open-source datasets. A primary focus is exploring the potential relationship between phonocardiography heart sounds and stenosed vascular sounds employing advanced classification techniques. Our approach denotes a novel contribution to the field, aiming to revolutionize the early detection and management of atherosclerosis.

1.4 The Outline of the Thesis

There are five chapters in the study. Chapter 1 presents the problem definition, covering atherosclerosis, diagnostic methods, and the study's acoustic principles. In addition, a literature survey and our approach will be presented. Chapter 2 introduces and elaborates on feature extraction and classification methods employed in this study. Chapter 3 focuses on heart sound classification, including the heart sound dataset, our implementation of feature extraction and classification techniques, and their different combinations. Chapter 4 explores the connection between heart and vascular sounds, covering information on vascular sound datasets. Chapter 5 offers conclusions and suggestions for future research.

CHAPTER 2

FEATURE EXTRACTION AND CLASSIFICATION METHODS

2.1 Feature Extraction Methods

2.1.1 MFCC Heat Maps

Mel-Frequency Cepstral Coefficients (MFCC) heat map is a type of representation used in audio and speech processing. It defines the spectral characteristics of sound and they are computed using the Mel Frequency Scale. It is a well-known and valuable method to extract features from an audio signal useful for various applications, such as speech recognition and sound classification.

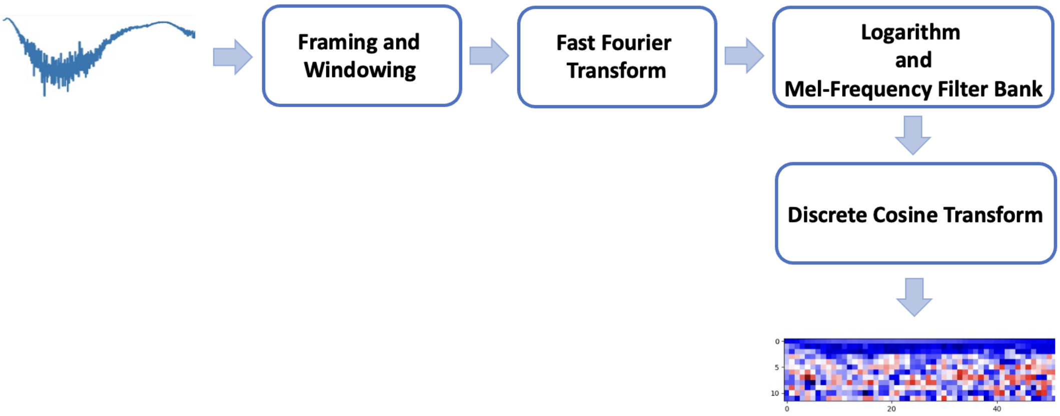

Framing and Windowing
Fast Fourier Transform
Logarithm and Mel-Frequency Filter Bank
Discrete Cosine Transform
0
5
10
0
20
40

Figure 2.1: Visualization of MFCC Calculation Process

Calculating MFCCs requires following a series of steps. The first step is dividing the audio signals into smaller frames. The amplitude spectrum of each frame is obtained by utilizing Fast Fourier Transform (FFT) to translate the signal into the frequency domain. Subsequently, the logarithm operation of the spectrum follows the process and the spectrum is transformed into a Mel-scale representation. Finally, The discrete cosine transform is implemented to extract distinctive frequency features [38]. Figure 2.1 is the visual representation of MFCC calculation process.

2.1.2 OpenL3

OpenL3 algorithm was originated from the L^3-Net. Arandjelovic and Zisserman [34] developed the L^3-Net (Look, Listen, and Learn Net), which integrates audio-visual correspondence. Three key goals drove their research. The first motivation behind the research is learning from generous and free resources; using widely available videos provides both visual and audio data. The next one is mimicking infant learning patterns, similar to how infants develop their visual and auditory skills by observing and listening to the world around them. The third objective is to evaluate the performance of the networks for different tasks.

The core objective was to create a system that independently learns visual and auditory semantic information by watching and listening to numerous unlabeled videos. This was accomplished through a novel learning task named Audio-Visual Correspondence (AVC), which trains both visual and audio networks from the beginning. An illustration in Figure 2.2 shows this audio-visual correspondence method.

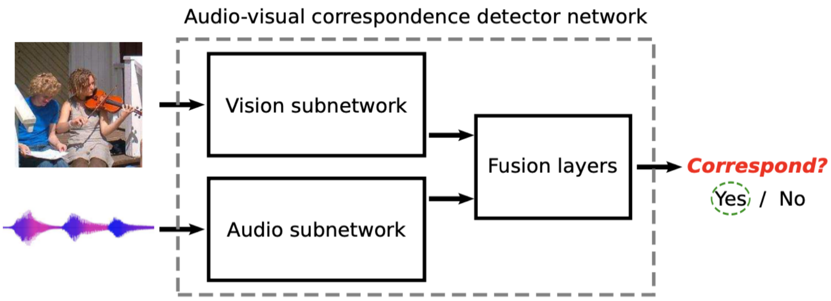


Figure 2.2: Illustration of Audio-Visual Correspondence Method retrieved from [34]

In the network architecture,shown in Figure 2.3, it is divided into audio and vision networks. While the image input size is $224 \times 224 \times 3$, the audio signal is utilized as a $257 \times 199 \times 1$ sized log-spectrogram, which belongs to 1-second and 48 kHz audio. There are four convolutional and max-poling layers, one after another, in both subnetworks. Also, each convolution has batch normalization and ReLU activation functions. After the outputs of visual and audio subnetworks are concatenated, it follows the fully connected layer, ReLU activation function, fully connected layer, and finally, softmax activation function. They utilize a max-pooling technique that results in a final embedding dimensionality of 6144.

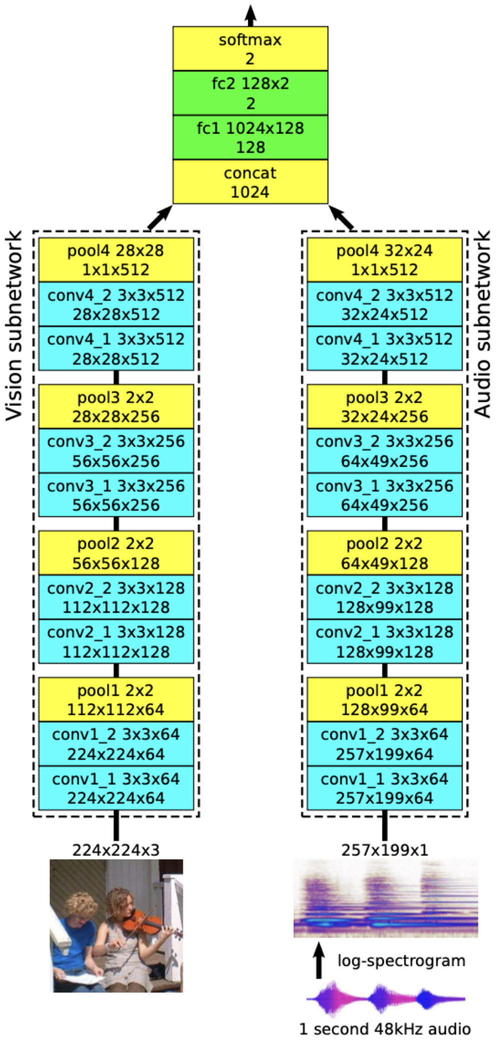


Figure 2.3: L^3-Net Arhitecture retrieved from [34]

Although the embedding shows potential for future applications, certain design decisions that could affect its effectiveness and computational efficiency remain unclear. To gain a deeper insight into how the embedding functions, Cramer et al. [39] explored three different design approaches.

The first one is the input representation of audio. They explain the first approach as the original L3-Net utilizes a spectrogram based on the linear frequency and logarithmic magnitude for its audio subnetwork input. However, Mel-frequency logarithmic magnitude spectrograms are more commonly used in machine learning applications. These Mel spectrograms capture perceptually significant and require fewer frequency bands than linear spectrograms. Notably, the Mel scale's quasi-logarithmic frequency allows for better pattern consistency in pitch-shifted sounds and makes it more effective for convolutional filters.

Figure 2.4 illustrates the study's findings. It shows that Mel spectrograms consistently outperform linear spectrograms across all datasets. The 256-bin Mel spectrogram yields the best results, significantly outperforming others in UrbanSound8K and ESC-50 datasets. This suggests that Mel spectrograms are more efficient at capturing crucial audio information. However, in the small DCASE 2013 SCD dataset, all types of embeddings show similar high accuracy.

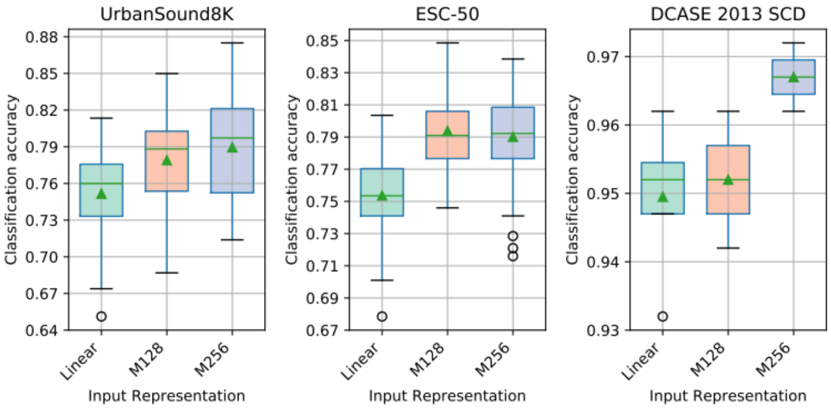


Figure 2.4: Classification Accuracy Against Input Representation retrieved from [39]

The second one is training domain and downstream tasks. The L^3-Net creators trained their model using videos rich in Audio-Visual Correspondence (AVC). The initial attempt was to use the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M), provided by [40]. It is the most extensive public multimedia collection ever made available.

This dataset comprises 100 million media items, including around 99.2 million photographs and 0.8 million videos, all under a Creative Commons license. Each item is accompanied by various metadata details, such as the Flickr ID, owner's name, camera used, title, tags, location, and source of the media. This dataset offers an exhaustive overview of the evolution of photo and video capture, tagging, and sharing, covering the period from Flickr's start in 2004 to early 2014.

Then, the AudioSet [41] was initialized. It consists of 2,084,320 YouTube videos and 527 labels accordingly. There is a vast array of 10-second audio clips and hand-labeled videos. The compilation of this dataset involved human annotators confirming the sounds in these YouTube segments. These segments were selected for annotation based on YouTube's metadata and content-driven search methods.

AudioSet's labels shed light on the video content and its influence on the model. They focused on videos featuring musical instrument performances, while their intended applications dealt with environmental sounds.

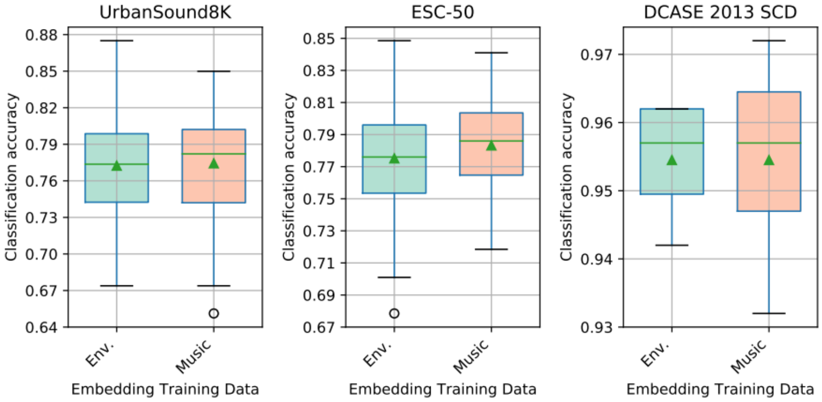


Figure 2.5: Classification Accuracy Against Different Training Datasets retrieved from [39]

This study explores whether aligning the training audio with the task-specific audio improves results, expecting such matching to enhance performance. Cramer et al. [39] investigated the impact of this domain alignment on the performance of the primary classification task, illustrated in Figure 2.5.

Contrary to expectations, aligning the audio domains did not yield a positive effect on performance. Specifically, in the context of the ESC-50 task, a slight decline in performance was observed. This outcome suggests the potential superiority of selecting audio content based on its capacity to enhance the discriminative ability of the embedding, irrespective of the audio domain of the subsequent task. In this regard, videos featuring musical instrument performance are presumed to exhibit a higher degree of AVC than environmental videos, which may be a more critical factor in evaluating the efficacy of the resultant embedding.

The third design concern is the quantity of training data. Arandjelovic and Zisserman [34] trained their models using 60 million samples. They did not address how the data volume impacts the embeddings' effectiveness. Given the considerable time and computational resources required for training, assessing the balance between the quantity of training data and the performance in subsequent classification tasks is valuable.

Figure 2.6 presents the results for UrbanSound8K and ESC-50 in the top and bottom plots, respectively. In UrbanSound8K, accuracy improvements start to plateau after training with 13 million samples, achieving roughly 77% accuracy. For ESC-50, a similar trend is observed, with diminishing returns after using 40 million samples, reaching around 79% accuracy. These findings imply that for training under limited resources, using at least 40 million samples is advisable for optimal training of the L^3-Net embedding [39].

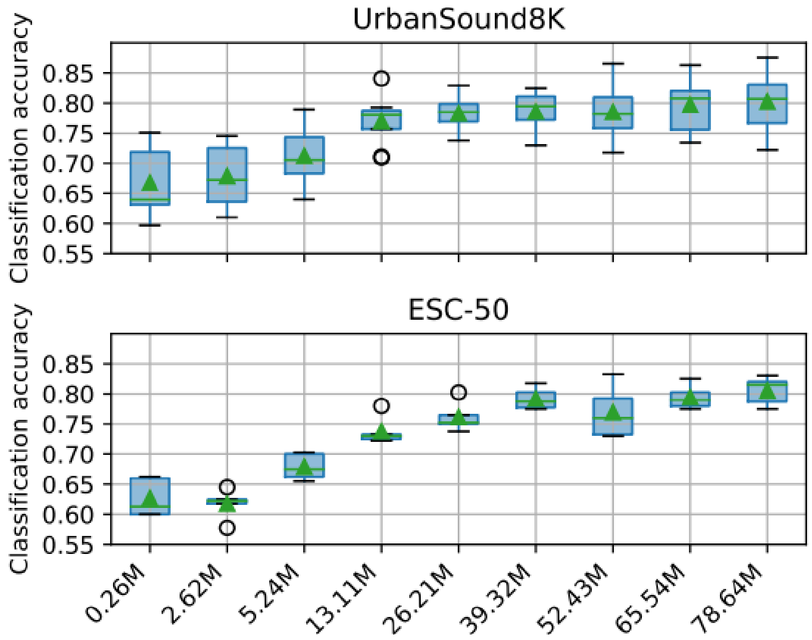


Figure 2.6: Comparison of Classification Accuracy Relative to the Number of Training Samples in Embedding Model Training retrieved from [39]

2.1.3 VGGish

The origin of the VGGish model is the VGG architecture for image recognition. The Visual Geometry Group created the VGG network. Simonyan and Zisserman [42] introduced a deep CNN for large-scale image recognition. The VGG architecture became famous for its depth and performance in the ImageNet competition. The adaptation of VGG for audio was developed by Hersley et al. [35], researchers at Google, and VGGish is part of their AudioSet project.

Image classification has seen considerable progress due to the introduction of extensive datasets like ImageNet and the application of CNN architectures, including AlexNet, VGG, Inception, and ResNet. The researchers mentioned that they had been encouraged by these developments and then investigated whether large datasets and CNNs could also improve performance in audio classification tasks.

Gemmeke et al. [43] created AudioSet, sourced from YouTube videos, to narrow the gap in data availability between image and audio research domains. This initiative marked a significant development in the audio domain, with the dataset undergoing continuous improvements and expansions. VGGish, leveraging this initiative, used a massive YouTube dataset comprising 100 million YouTube videos, including 70 million training videos, 10 million evaluation videos, and 20 million videos that we used for validation. Each video in the training program has an average duration of 4.6 minutes. In total, there are 5.4 million hours of training available. Each video is labeled automatically from approximately 30,000 labels based on a combination of metadata.

The audio from the dataset is segmented into 960 ms frames, which inherit the labels of their respective videos. It results in around 20 billion samples. They process each frame using Fourier transform with 25 ms windows every 10 ms. The spectrograms are produced and then converted into 64 mel-frequency bins. This process generates log-mel spectrograms that serve as inputs for all classifiers. They created mini-batches of 128 examples by randomly selecting from these patches for training.

The experiments are conducted with TensorFlow utilization with multiple GPUs asynchronously and the Adam optimizer. The VGGish Network architecture is depicted in Figure 2.7. After each convolutional layer, batch normalization was applied. For the final layer, sigmoid was chosen rather than softmax in case of multiple labels. The loss function used was cross-entropy. Since there is no overfitting, there is no regularization technique.

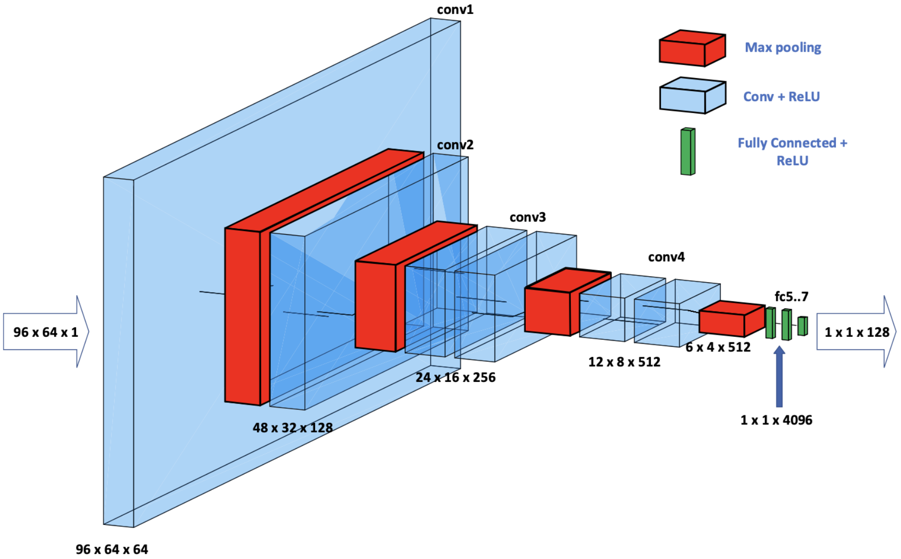


Figure 2.7: VGGish Network Architecture retrieved from [44]

The team established a fully connected deep neural network (DNN) as the baseline, using approximately 30,000 labels for training and evaluation. They optimized GPUs and learning rates to maximize frame-level classification accuracy. The best-performing baseline model has 3 layers with 1000 units each, a learning rate 3×10^{-5} by means of 10 GPUs and 5 parameter servers, and about 11.2 million weights.

For AlexNet, modifications included adjusting the stride in the initial convolutional layer and replacing local response normalization (LRN) with batch normalization due to different input sizes. The final layer was also altered, which led to a total of 37.3 million weights. Unlike the original AlexNet was trained with 20 GPUs and 10 parameter servers, without distributing filters across devices.

In adapting VGG, the changes were limited to the final layer and batch normalization instead of LRN. The audio version of VGG had 62 million weights. Adjusting initial strides was tested, but maintaining the original stride proved more effective. Training used 10 GPUs and 5 parameter servers.

For InceptionV3, the team modified the initial layers to the MaxPool and removed the auxiliary network, leading to an audio variant with 28 million weights. Adjustments in the Average Pool size were made to suit audio activations better. This model was trained with 40 GPUs and 20 parameter servers.

Lastly, ResNet-50 was modified by removing the stride from the initial convolution and adjusting the Average Pool size. This resulted in an audio-specific version with 30 million weights. It was trained using 20 GPUs and 10 parameter servers.

Each of these models represents a tailored adaptation of a successful image classification architecture, reconfigured to address audio data's specific challenges and characteristics.

2.2 Classification Methods

2.2.1 Convolutional Neural Network

Artificial Neural Networks (ANNs) are computational models directly based on the functioning of the human brain's nervous systems. These networks are composed of numerous interconnected processing units, also called neurons. The neurons work together in a coordinated and distributed manner to process input data effectively. The illustration of the analogy between the real and artificial neuron is given in Figure 2.8.

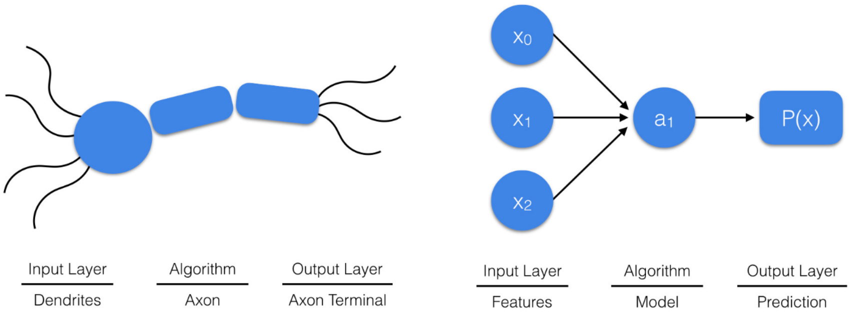


Figure 2.8: Analogy between the Human and Artificial Neuron Retrieved from [45]

Figure 2.9 shows the fundamental structure of an ANN. The process begins by feeding a multidimensional vector into the input layer and then forwarding it to the hidden layers. The hidden layers of the ANN are responsible for evaluating and processing the information received from the previous layer. Therefore, they assess how random variations within themselves negatively or positively impact the final output. This phase is known as the learning process. This structure is typically called deep learning if several hidden layers layered on top of each other.

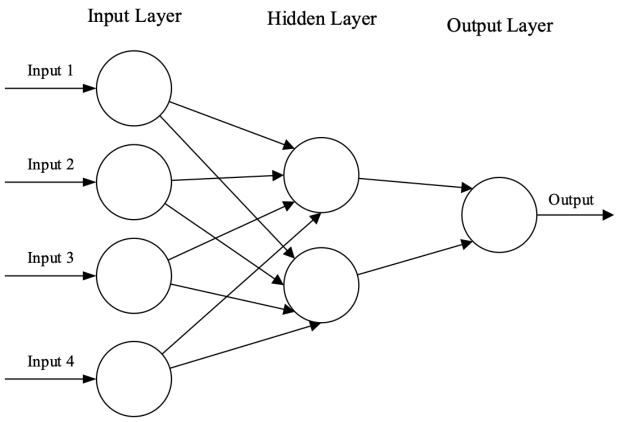


Figure 2.9: Simple Neural Network Representation retrieved from [46]

Although the mathematics dates back earlier, the convolutional neural networks introduced by LeCunn et al. [47] in 1989. They used the backpropagation algorithm for a neural network (NN) architecture for recognizing handwritten digits in images. This network, named LeNet, is widely regarded as the first successful implementation of a convolutional neural network. It established the foundation for contemporary CNN architectures.

CNNs resemble ANNs because they consist of neurons that have self-improvement ability through learning. The neurons process inputs by executing mathematical operations, like ANNs. The network conveys the weights from the initial raw vectors to the output. This structure enables CNNs to handle tasks such as image and video recognition efficiently by reducing parameters and computations compared to traditional ANNs.

The architecture of CNNs could be described as a series of distinct layers. These layers have specific roles in the pipeline. Convolutional layers extract spatial hierarchies of features from the input by empowering the hyperparameters. Activation layers introduce non-linearities and allow the network to learn complex patterns. Pooling layers reduce the data's dimensionality and enhance the computational efficiency and feature robustness. Fully connected layers then interpret these features to make final predictions or classifications. Dropout layers prevent overfitting and ensure the model's evaluation performance for unseen data.

Key hyper-parameters such as kernel size, number of filters, stride, and padding shape the convolutional layers structure. These parameters determine how the network filters and processes input data while capturing essential features. Similarly, learning rate, epochs, and batch size control the overall learning process and ensure the efficiency of the network.

In this context, a significant application of CNNs in the field of medical diagnostics can be seen in the work of Rubin and his teammates [31]. They proposed an algorithm that integrates CNN and MFCC heat maps. After the signals' arrangement, ninety thousand MFCC heat maps were used as training and validation sets as input.

Figure 2.10 shows the architecture of the convolutional neural network to predict normal versus abnormal sounds. The audio files undergo pre-processing using MATLAB, facilitated by pre-processing codes offered by PhysioNet at the challenge. For network implementation, TensorFlow is employed as the framework of choice. While the source code for the inference model is accessible, it is noteworthy that the source code for the training model is not provided.

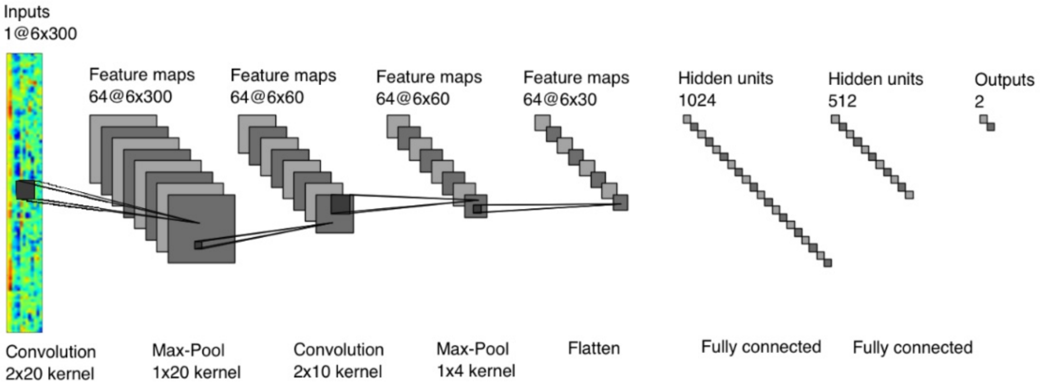


Figure 2.10: Convolutional Neural Network retrieved from [31]

The hyper-parameters and network parameters are detailed in Table 2.1 and Table 2.2.

Table 2.1: Hyperparameter Values

Hyperparameter	Value
Learning Rate	0.00015822
Beta	0.000076253698849
Dropout	0.85565561

Table 2.2: Network Parameter Values

Network Parameter	Value
Regularization Type	L_2
Batch Size	256
Optimization	Adam

The final test scoring of the network, conducted after the completion of training, are 76.5% sensitivity, 93.1% specificity and 84.8% overall in the challenge.

2.2.2 Support Vector Machine

Support vector machines, abbreviated as SVMs, is a supervised learning method. It was initially created by Vapnik et al. in the 1960s to recognize patterns and classify data, which was part of their more comprehensive research on machine learning and decision-making theory. Later, in 1995, Vapnik and Cortes [48] published an important paper that detailed SVM more thoroughly.

They first mentioned the SVMs as Support-Vector Network and introduced them as a new learning method for two-class classification. The algorithm is not only reasonable for classifying data but also applicable for regression tasks.

There are three concepts to understand the algorithm background. The primary concept of SVMs is identifying the optimal decision boundary, called a hyperplane, separating different data classes. This boundary is selected to maximize the margin, which is defined as the space between the hyperplane and the closest points from each class. Support vectors are the essential data points that lie nearest to the decision boundary in an SVM model, critically influencing the model's classification decision. Figure 2.11 presents visual representations of the concepts.

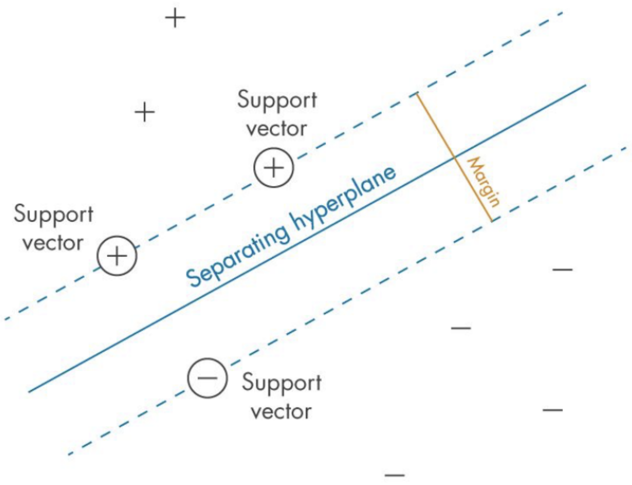


Figure 2.11: Visual Representation of SVM Basic Concepts retrieved from [49]

SVMs rely on a set of mathematical functions called kernels. The kernel function transforms input data into the desired form. There are three common kernels.

The linear kernel is a straightforward function that calculates the value through the inner product of two vectors. The dot product shows the similarity between these vectors. This simplicity is a significant advantage, particularly in cases where the data points exhibit a linear relationship and large-scale applications. The mathematical representation of the linear kernel is given by:

$$K(x, y) = x \cdot y \tag{2.1}$$

where x and y stand for the feature vector of the data points.

The polynomial kernel is a more complex mathematical function than the linear kernel. It plays a critical role when coping with nonlinear data by mapping the original input features into a specified dimensional space. This transformation is achieved through the use of polynomials of the original variables. Doing so facilitates the SVM's ability to learn nonlinear models that would be challenging or impossible to fit using a linear kernel. The mathematical representation is expressed as follows:

$$K(x, y) = (\gamma \cdot x \cdot y + c)^d \tag{2.2}$$

where γ stands for a scale factor, c is a constant, and d term shows a polynomial degree.

The Radial Basis Function (RBF) kernel is a common choice for nonlinear data. This function transforms samples into a higher dimensional space using the Gaussian function. It is a robust and effective function when the data distribution is unknown or exhibits variability. This versatility makes it appropriate for a wide range of applications. The mathematical expression is as follows:

$$K(x, y) = e^{(-\gamma \|x-y\|^2)} \tag{2.3}$$

where γ is the Gaussian distribution parameter.

2.2.3 K-Nearest Neighbor

K-Nearest Neighbors, abbreviated as KNN, is a versatile distance-based supervised learning method. The conceptual foundations of KNN can be traced back to the work of Fix and Hodges [50] in 1951. Later, Cover [51] further developed these ideas into today's KNN for classification.

Today, KNN is recognized as an essential tool in machine learning. Unlike many contemporary machine learning methods, KNN stands out for its simplicity and intuitive approach to classification and regression tasks.

The core idea of KNN depends on the concept that similar samples tend to be found close to each other. A distance metric defines this proximity, which computes the similarity between instances. There are commonly used distance metrics. Euclidean distance is a very well-known measure. It measures the distance between two vectors and is restricted to samples with real values. The Euclidean distance formula is given by:

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.4}$$

where x and y represent the first and second data points with n dimensions.

Another widely used metric is the Manhattan distance, commonly known as taxicab or city block distance. It calculates the sum of the absolute differences of their coordinates and is more suitable than the Euclidean distance for grid-like data structures. The formula of Manhattan distance is given by:

$$D(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.5}$$

where x and y represent the first and second data points with n dimensions.

The Minkowski is the generalization of Euclidean and Manhattan distances. It involves a parameter p, which leads to different distance measures. Specifically, setting p to two yields the Euclidean distance, while a p-value of one corresponds to the Manhattan distance. The mathematical representation of Minkowski distance is given by:

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{2.6}$$

where x and y represent the first and second data points, p is a parameter that determines the type of distance.

The algorithm involves identifying the 'k' nearest neighbors of a given data point and making predictions based on these neighbors. The most common class among the neighbors is assigned to the data point for classification, while the average of the neighbors' values is used for regression. The idea behind the KNN could be illustrated as in Figure 2.12.

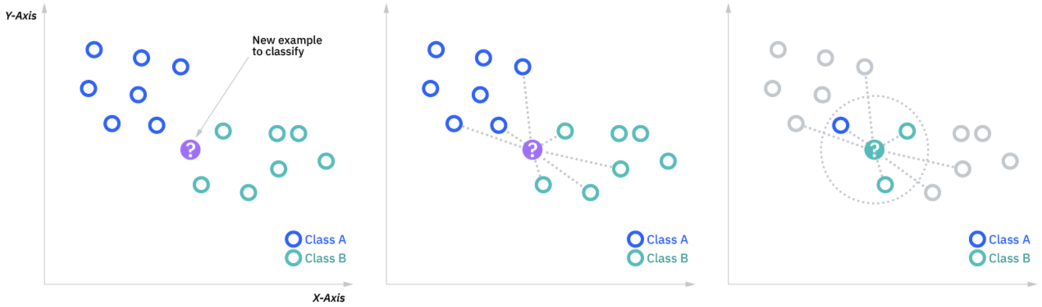


Figure 2.12: Illustration of KNN Classification Approach retrieved from [52]

This chapter presents key findings from relevant studies and experiments, providing a comprehensive understanding of the methodologies employed in extracting features and classifying audio signals. In the subsequent chapters, these methods will be further utilized and evaluated in the context of specific heart and vascular acoustic analysis.

CHAPTER 3

CLASSIFICATION OF HEART SOUNDS

3.1 Dataset (PhysioNet Heart Sound Database)

The foundation of our exploration lies in the rich and diverse dataset known as the PhysioNet Heart Sound Database, a publicly accessible repository curated through electronic stethoscope recordings for the Computing in Cardiology Challenge [53]. The collection comprises nine distinct databases of heart sounds. They were assembled by different research groups from seven nations across three continents over ten years. The details about the databases are described as following:

MIT heart sound database ,created by Professor John Guttag, Dr. Zeeshan Syed, and their team, contains 409 heart sound recordings from 121 people. These were recorded using a sophisticated electronic stethoscope and an ECG, with a high-quality sampling rate of 44.1 kHz and 16-bit clarity.

The database groups the subjects into two categories: healthy individuals and patients those with mitral valve prolapse murmurs, with harmless murmurs, with aortic disease, and with other heart conditions. These diagnoses were all double-checked with echocardiograms at the Massachusetts General Hospital.

Recordings are between 9 to 37 seconds. They were captured in various places, like homes and hospitals, and included background noises like talking. Despite this, the recordings are still very useful for studying heart sounds.

AAD heart sounds database was created with contributions from Schmidt and his team from Aalborg University. It includes heart sound recordings taken from 151 patients at the Cardiology Department in Aalborg Hospital, Denmark, using a Littmann

E4000 electronic stethoscope with 4 kHz sampling rate and 16 bit quantization. These patients were being evaluated for (CAD) via coronary angiography.

For this database, ’normal’ and ’abnormal’ classifications were based on the presence of heart valve defects, either noted in medical records or detected through clear heart murmurs. Each patient provided one to six PCG recordings, accumulating to 695. Most of these recordings were standardized to 8 seconds in length.

AUTH heart sounds database was created with contributions from Papadaniil and Hadjileontiadis from The Aristotle University of Thessaloniki in 2014. It consists of 45 individuals’ heart sounds captured at the Papanikolaou General Hospital in Thessaloniki, Greece. The age range is from 18 to 90 years. The recordings used a custom electronic stethoscope named AUDIOSCOPE which has 4 kHz sampling rate and 16-bit depth. Subjects were divided into three groups: 11 with normal heart sounds, 17 with aortic stenosis, and 17 with mitral regurgitation, all diagnosed by echocardiogram. Heart murmurs were recorded at the chest location where they sounded clearest, while normal heart sounds were recorded at the apex of the heart. Durations range from 10 to 122 seconds.

TUT heart sounds database,provided by Naseri and Homaeinezhad from Toosi University of Technology in 2013, contains heart sound data from 28 individuals without heart conditions and 16 patients with diagnosed valve diseases, confirmed through echocardiographic evaluation. Recordings were captured utilizing a high-tech 3M Littmann 3200 electronic stethoscope at four distinct heart areas: the pulmonic, aortic, tricuspid regions, and the apex. These sounds were digitally captured with a 4 kHz sampling rate and a precision of 16 bits for 15 seconds per recording. In total, the database holds 174 PCG recordings, with two participants providing three recordings each.

UHA heart sounds database, provided by Moukadem et al. from The University of Haute Alsace’s , includes 79 recordings featuring both normal and abnormal heart sounds captured with high-quality stethoscopes at an 8 kHz sampling rate. The collection is split into two groups for the normal sounds: one from 19 healthy individuals and another from six astronauts involved in a space simulation project. The recordings vary in length but average around 14 seconds for the healthy group and 10 seconds for the astronauts.

The database also contains recordings from 30 hospital patients, ranging from 44 to 90 years old, some of whom were recorded pre- and post-heart valve surgery. A cardiologist diagnosed their conditions using ECG and echocardiography. Some patients had artificial valves, while others had rhythm problems associated with heart disease. These recordings last between 6 and 49 seconds.

DLUT heart sounds database was provided the courtesy of Tang and Li's research at The Dalian University of Technology, featuring 174 healthy individuals and 335 coronary artery disease (CAD) patients. Healthy subjects were predominantly male, young adults, while CAD patients spanned a broader age range, averaging 60 years old. Heart sounds from patients was collected from the chest's mitral area using a standard electronic stethoscope with 8 kHz sampling rate and 16-bit resolution at a medical facility. Healthy participants were recorded at the university's lab using various sensors. Expert cardiologists confirmed CAD diagnosis. Recordings vary from 3 to 98 seconds.

SUA heart sounds database,assembled by Samieinasab and Sameni at The Shiraz University, consists of audio recordings from 112 individuals, 79 healthy subjects, and 33 patients with heart conditions. The age band ranges from 16 to 88 years. For capturing the heart sounds, the JABES electronic stethoscope was utilized, mainly positioned above the heart's apex region, while recordings were managed using Audacity software. Subjects were recorded once, save for one healthy individual who contributed three recordings, yielding 81 normal and 33 abnormal heart sound files. Recordings lasted between 30 to 60 seconds, with a standard sampling rate of 8 kHz and 16-bit depth; however, a few recordings were captured at much higher rates.

SSH heart sounds database contains 35 heart sound recordings, which were collected from patients at the Skejby Sygehus Hospital in Denmark. This collection includes contributions from 12 individuals with no heart issues and 23 patients diagnosed with heart valve defects. Each heart sound was recorded at the second intercostal space near the right side of the sternum. The duration of the recordings ranges from about 15 seconds to 69 seconds and all sounds were recorded with 8 kHz sampling rate.

This diverse assembly process has led to significant variations in the recording equipment, the anatomical sites of recording, the quality of the data captured, and the types of patients. It includes a total of 2435 recordings of heart sounds. They were obtained from 1297 individuals, containing both healthy subjects and those with various cardiac conditions but typically they suffer from heart valve defects and CAD. These recordings were acquired from multiple settings, ranging from clinical to nonclinical scenarios like home visits, utilizing different types of equipment.

The heart sound recordings were captured from various points on the body but the focus was on the 4 conventional sites. Optimal heart sound recordings are obtained from these designated areas on the thorax, namely the aortic, pulmonic, tricuspid, and mitral areas. Figure 3.1 illustrates these specified regions, along with the anatomical landmarks of the right ventricle (RV) and left ventricle (LV), which are crucial for understanding the orientation of the recordings. The aortic region (AO) is centered in the second intercostal space to the right (1), where the aortic valve sounds are best heard. The pulmonic area (PA) is situated in the second intercostal space along the left border of the sternum (2), a prime location for capturing the sounds of the pulmonic valve. The tricuspid region (3), found in the fourth intercostal space adjacent to the left side of the sternum, allows for the auscultation of the tricuspid valve sounds. Finally, the mitral region (4) is placed at the heart's apex, specifically in the fifth intercostal space that aligns with the midclavicular line, which is ideal for listening to the mitral valve sounds [54].

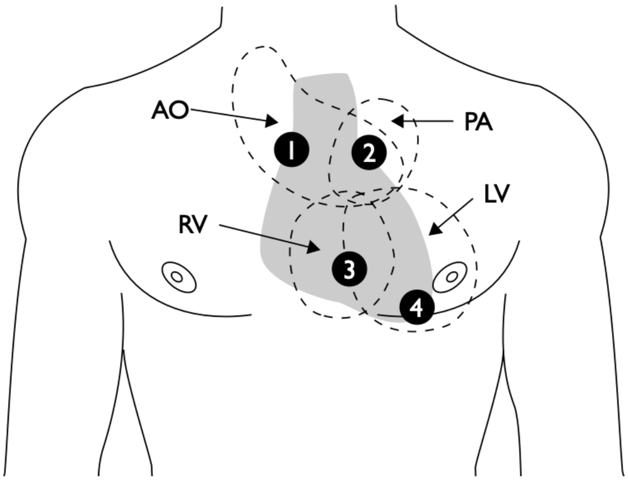


Figure 3.1: Optimal Auscultation Points retrieved from [54]

The duration of recordings ranges from a few seconds to several minutes. Each recording file starts with an identical letter followed by a sequentially assigned yet random numerical value in every dataset. Files from a single patient are unlikely to be in consecutive numerical order. The training and test datasets are established as two distinct groups that do not overlap and they have unbalanced distribution. The duration of the recordings ranges from a few seconds to several minutes. All recordings have been converted to .wav format at a sampling rate of 2,000 Hz.

Since the combination of all the datasets shows poor performance, four specific heart sound databases, the MIT, the AAD, the AUTH, and the TUT Heart Sound Databases, are employed for these experiments. These databases were selected for their robustness and diversity regarding heart sound recordings.

The distribution of normal and abnormal heart sounds within these databases is critical to this research. This distribution is detailed in Table 3.1, which illustrates the balance or imbalance between normal and abnormal heart sounds in both datasets.

Table 3.1: Distribution of Normal and Abnormal Sizes Across Different Categories

Category	**Total Size**	**Abnormal Size**	**Normal Size**
A-Dataset	409	292	117
B-Dataset	490	104	386
C-Dataset	31	24	7
D-Dataset	55	28	27
E-Dataset	2141	183	1958
F-Dataset	114	34	80
Total	**3240**	**665**	**2575**

3.2 Environment

The implementation of the entire study was conducted within the Google Colab environment using Python. The dataset, initially uploaded to Google Drive, was subsequently mounted onto Colab, and the dataset's path was established. This step facilitated the data handling process.

Additionally, PyTorch was employed to design and train the neural networks. Its user-friendly interface and powerful tools for neural network construction and training greatly simplified the process. This streamlined approach of managing data in Colab and leveraging PyTorch's capabilities supported a smooth and effective implementation of the study.

3.3 Prediction and Performance Metrics Definition

In this part, the definitions of sensitivity, specificity, and overall accuracy is given by Equations 3.1, 3.2, 3.3.

Sensitivity, as represented in Equation 3.1, is a measure of the actual positives, which shows the correctly identified disease by the model in this context.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.1}$$

Specificity, as shown in Equation 3.2, is an evaluation metric of the actual negatives that are correctly identified in healthy instances.

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{3.2}$$

Accuracy, as expressed in Equation 3.3, is defined as the average of specificity and sensitivity values in this particular application.

$$Accuracy = \frac{Sensitivity + Specificity}{2} \tag{3.3}$$

3.4 Algorithm Implementations

3.4.1 MFCC Heat Maps

The challenge organizer [53] provides the pre-processing part. The software provides unique audio filtering, normalization, and segmentation functions. Therefore, the audio files were pre-processed in MATLAB with provided codes. MATLAB gets .wav files as 2 kHz and decreases it to 1 kHz. The illustration is given in Figure 3.2. As a result, the set of functions created 2575 normal and 665 abnormal .wav files.

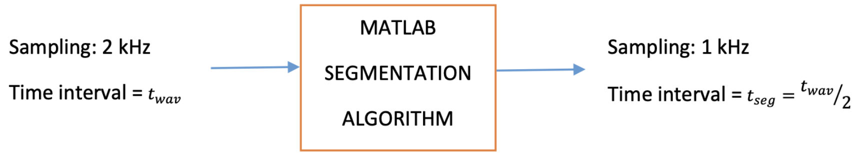

Sampling: 2 kHz
Time interval = t_{wav}
MATLAB
SEGMENTATION
ALGORITHM
Sampling: 1 kHz
Time interval = $t_{seg} = {}^{t_{wav}}/_{2}$

Figure 3.2: Illustration of Sample Rate Decrease by Segmentation Algorithm

Obtaining Mel-Frequency Cepstral Coefficients (MFCCs) begins with acquiring .wav audio files. Segmentation is performed to facilitate the sampling issues after creating paths for both normal and abnormal records. The MFCC extraction operates on .wav files with a MATLAB output with a sampling rate of 2 kHz downsampled to 1 kHz [55]. Also, the window step is set to 0.01 seconds to get 10 ms intervals. The duration of signals is determined as 3 seconds. The summary of the process is illustrated in Figure 3.3. As a result, 90. samples created from 3340 records.

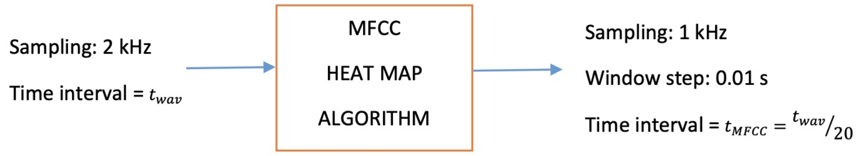

Sampling: 2 kHz
Time interval = t_{wav}
MFCC
HEAT MAP
ALGORITHM
Sampling: 1 kHz
Window step: 0.01 s
Time interval = $t_{MFCC} = {}^{t_{wav}}/_{20}$

Figure 3.3: Illustration of MFCC Heat Map Process Summary

MFCC coefficients provide essential frequency information enabling the identification of patterns in audio signals. The coefficients are then transformed into informative heat maps, which display the distribution of frequency features across audio frames. The illustration MFCC heat map Process summary is given in Figure 3.3.

The waveform representation of the initial normal sample is depicted in Figure 3.4, while the corresponding MFCC heat map for the same sample is presented in Figure 3.5.

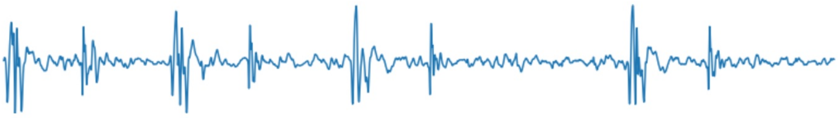


Figure 3.4: Normal(Healthy) Audio Sample Waveform

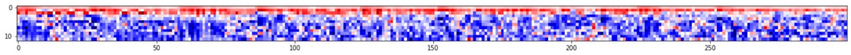


Figure 3.5: Normal Sample MFCC Heat Map

The waveform representation of the initial abnormal sample is displayed in Figure 3.6, along with the corresponding MFCC heat map, which is presented in Figure 3.7.

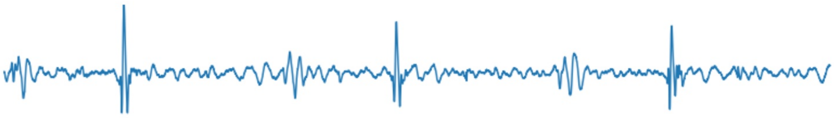


Figure 3.6: Abnormal Audio Sample Waveform

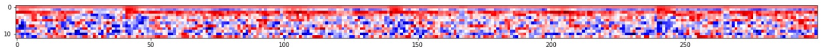


Figure 3.7: Abnormal Sample MFCC Heat Map

3.4.2 OpenL3

The usage of the OpenL3 algorithm is explained detailed in [56]. Figure 3.8 illustrates the implementation steps. In the development of the feature extraction process, the code initiates by importing essential libraries: 'openl3' for sophisticated audio feature extraction, 'soundfile,' and 'scipy.io.wavfile' for efficient audio file manipulation. This sets the foundation for subsequent operations. Three pivotal lists—'trainingLabels,' 'trainingFeatures,' and 'trainingRecords'—are then established to systematically store the labels, extracted features, and audio records.

This organization facilitates streamlined processing and analysis. The procedure's core involves iterating over each audio record in 'recordsList.'

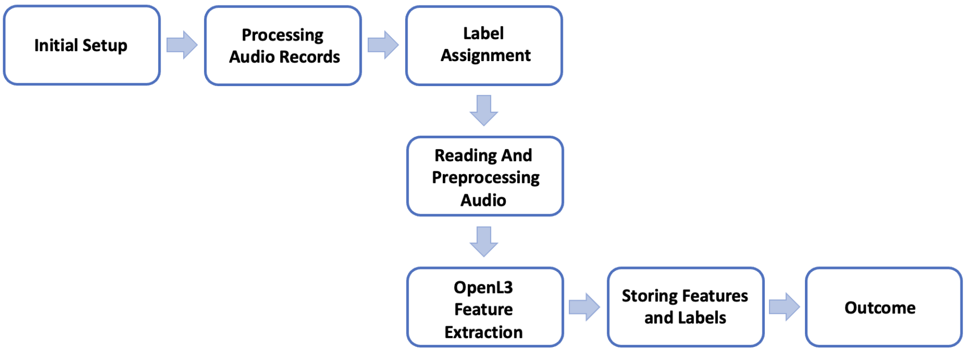

Initial Setup
Processing Audio Records
Label Assignment
Reading And Preprocessing Audio
OpenL3 Feature Extraction
Storing Features and Labels
Outcome

Figure 3.8: Illustration of OpenL3 Algorithm Implementation

Label assignment is a critical step where each record is categorized, with the label '1' assigned if it exists in a temporary normal list 'recordsListTmp'; otherwise, it is labeled '0'. This binary classification is essential in the subsequent analysis.The sophistication of the process is further evidenced in the audio reading and preprocessing phase.

Each audio signal is read from a .wav file and transformed into a floating-point format to ensure uniformity and compatibility with the feature extraction tools. A key aspect here is the standardization of the sampling rate to 48 kHz, a prerequisite for the 'openl3' library, achieved through resampling if the original rate deviates from this standard. Following this, the 'openl3' library is employed to extract audio features, known as embeddings, from the processed audio signal. These embeddings, encapsulated in the variable 'emb' are of a fixed size of 512, aligning with the parameters of the feature extraction algorithm.

The extracted features and corresponding labels are then meticulously appended to the 'trainingFeatures' and 'trainingLabels' lists. This systematic collection of data is essential for the forthcoming analytical stages. Each step of the process is marked by an output message, indicating the successful processing of each record.

Conclusively, the procedure ensures that the 'trainingFeatures' list comprehensively contains the features for each audio file, while the 'trainingLabels' list accurately represents their respective labels. This structured approach demonstrates a methodical and efficient audio feature extraction and labeling process, forming a crucial part of the data preparation phase in audio analysis research.

3.4.3 VGGish

The overarching steps in the implementation of the VGGish algorithm are summarized in Figure 3.9. The first section of the code involves installing necessary dependencies and setting up VGGish, a deep-learning model developed for audio analysis. This is achieved by installing ’TensorFlow’, a comprehensive machine-learning framework, and ’SoundFile’, a library dedicated to reading and writing sound files. Following this, the TensorFlow models repository is cloned from GitHub. Additional specific requirements for VGGish are then installed.

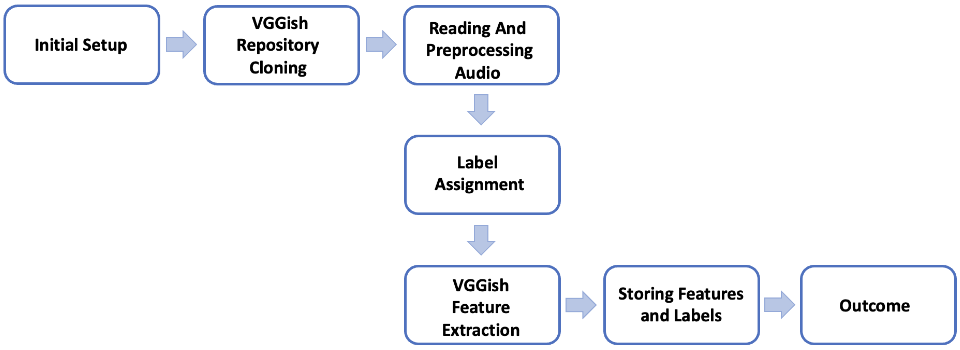

Initial Setup
VGGish Repository Cloning
Reading And Preprocessing Audio
Label Assignment
VGGish Feature Extraction
Storing Features and Labels
Outcome

Figure 3.9: Illustration of VGGish Algorithm Implementation

Subsequently, essential modules for operating with the VGGish model are imported: ’vggish-slim’, ’vggish-params’, and ’vggish-input’. TensorFlow version 1 is also imported.

The VGGish model checkpoint is downloaded. It contains pre-trained weights vital for audio data processing. This step is crucial for enabling the VGGish model to process audio inputs effectively.

The main segment of the code is dedicated to processing audio data in batches. A batch size is defined, which can be adjusted based on memory limitations, and the number of batches is calculated by dividing the total number of audio records by the batch size. Two lists, ’trainingFeatures’, and ’trainingLabels’, are initialized to store the extracted features and their corresponding labels.

Using TensorFlow’s computational graph and session, the VGGish model is defined and loaded with ’vggish-slim’. The input and output tensors are retrieved to facilitate feeding audio data into the model and extracting embeddings.

In each batch of audio records, file paths are determined based on their identifiers (a, b, c, d), and labels are set for each record. The audio files are then processed to extract features using the VGGish model. This involves converting ’.wav’ files into a format compatible with the model using ’vggish-input.wavfile-to-examples’. It is pursued by running the TensorFlow session to obtain the embeddings. These embeddings and their labels are appended to the ’trainingFeatures’ and ’trainingLabels’ lists.

In conclusion, ’trainingFeatures’ contains the VGGish embeddings for each audio file, and ’trainingLabels’ holds the corresponding labels. This represents the completion of the process, yielding processed audio data ready for subsequent analytical or machine-learning applications.

3.4.4 Convolutional Neural Network

Figure 3.10 depicts the steps involved in implementing the CNN algorithm. To start, the script imports the required libraries for both data processing and neural network construction. Libraries like ’PyTorch’, ’TensorFlow’, ’NumPy’, and ’Pandas’ are used for various computational tasks. If it is available, GPU availability is checked to determine whether GPU acceleration can be utilized.

Data for the neural network is taken from audio recordings stored on a Google Drive mount. The script processes these recordings to create the dataset. The recordings are read and segmented using Mel-Frequency Cepstral Coefficients (MFCC). These MFCC features form the basis of the training data. Each audio segment is labeled as normal or abnormal and creates a dataset for supervised learning.

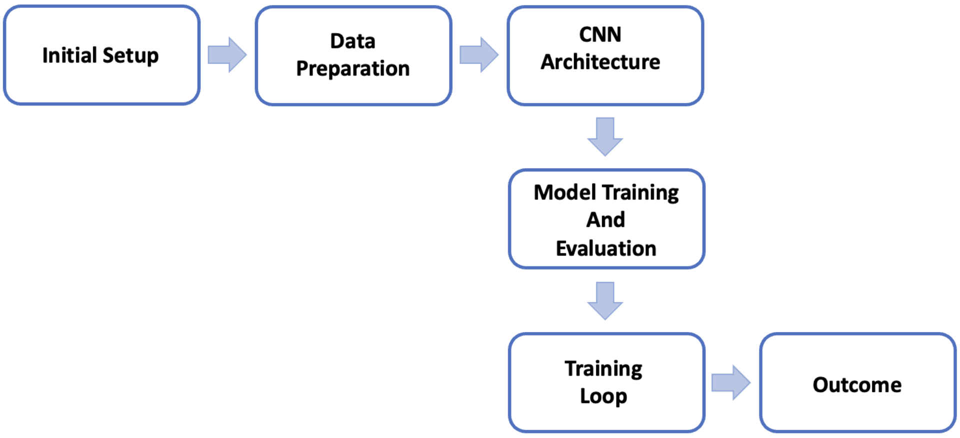


Figure 3.10: Illustration of CNN Algorithm Implementation

The neural network, a CNN, is defined with multiple layers, including convolutional layers, max-pooling layers, batch normalization, and fully connected layers. These layers are designed to process the input MFCC features effectively and capture relevant classification patterns.

Training the network involves feeding the data through the network, calculating loss using Cross-Entropy Loss, and adjusting the network's weights using Stochastic Gradient Descent (SGD) as the optimization algorithm. The network is trained over multiple epochs. Both training and validation datasets are calculated after each epoch. This process helps understand the model's performance and ensures it learns to classify the heart accurately sounds.

Additionally, functions are defined for visualizing data, resetting model weights, and checking the model's accuracy. These functions aid in model evaluation and understanding the network's learning process.

The script concludes with a training loop where the network undergoes training for a specified number of epochs, and its performance is evaluated in terms of accuracy on both training and validation datasets. The accuracy results are tracked to monitor the model's progress over time.

3.4.5 Support Vector Machine

The implementation steps of the SVMs algorithm is illustrated in Figure 3.11. The process begins with setting up the necessary environment for machine learning by installing 'scikit-learn'. It is a powerful library in Python known for its versatility in machine learning. This step is crucial as it provides the essential tools required for the task ahead. Following the installation, various modules are imported for different purposes: model selection, preprocessing, metric evaluation, and the SVM classifier.

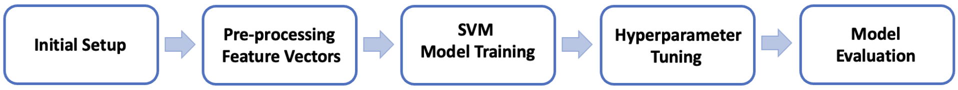


Figure 3.11: Illustration of SVM Implementation

In data preprocessing, the code first examines the length of the training features to ensure uniformity across the dataset. If the lengths vary, the sequences are padded to a uniform length using 'padsequences' from 'keras.preprocessing.sequence'. This standardization is crucial in preparing the data for effective feature extraction and subsequent processing.

The data then undergoes a transformation and scaling process. It is reshaped to match the input requirements of the SVM classifier, and scaling is performed using 'StandardScaler'.

Once the data is preprocessed, the focus shifts to training the SVM classifier. An SVM with a linear kernel is chosen and trained on the processed training data. The choice of the linear kernel is significant as it influences the classifier's decision boundary.

After training, the model's performance is evaluated on the test set using various metrics like accuracy, precision, recall, and F1-score. These metrics were obtained from functions like 'accuracyscore' and 'classificationreport'. They provide a comprehensive view of the model's effectiveness in making predictions.

In SVM, the regularization term C balances maximizing the margin and minimizing classification error. A smaller value of C results in a broader margin but allows more misclassifications, emphasizing the simplicity of the decision boundary. Conversely, a larger C value aims to minimize misclassifications, and it can lead to a more complex model with a narrower margin.

Therefore, while a larger C value in SVM can yield a model with fewer training errors, balancing this against the risk of overfitting is essential. This balance is crucial to ensure the model remains robust and performs well on the training and new data. The choice of C thus becomes a critical decision in SVM training, requiring careful tuning to hit the right balance between complexity and generalization.

The gamma parameter holds a significant influence on the RBF kernel. It determines the decision boundary's complexity. A high gamma value leads to more complex decision boundaries by paying closer attention to the training data. This increased sensitivity to individual data points allows the model to detect subtle and complex patterns within the training data. Nevertheless, this increased complexity comes with the risk of overfitting, especially in scenarios where the training data includes outliers.

On the other hand, a lower gamma value spreads the influence of each training example over a broader area and results in a more generalized decision boundary. This broader influence typically leads to more robust models that generalize better on unseen data. Yet, setting the gamma too low can lead to underfitting because the model becomes too simplistic and fails to capture critical patterns for accurate predictions. Therefore, finding the optimal gamma value is necessary in SVM model tuning.

Further refinement is achieved through hyperparameter tuning using 'GridSearchCV'. This process involves experimenting with different values of 'C', 'gamma', and kernel function to find the best combination. This step enhances the model's accuracy.

Finally, the best model from the grid search is used to make predictions on the test data. The evaluation of this model includes generating a confusion matrix and calculating sensitivity & specificity. These metrics offer deeper insights into the model's prediction.

3.4.6 K-Nearest Neighbor

Figure 3.12 provides a overhead illustration of the implementation steps for the K-Nearest Neighbors (KNN) method using the scikit-learn library. It begins with the importation of necessary modules: 'KNeighborsClassifier' for the KNN model, 'train-testsplit' for data segmentation, various metrics for performance evaluation, and 'StandardScaler' for feature normalization.

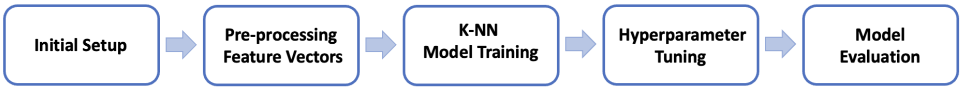


Figure 3.12: Illustration of KNN Implementation

The 3D input feature arrays ('trainingFeatures') are initially processed. This involves flattening or aggregating these features into a 2D format, executed by computing the mean along one axis and subsequently flattening the array. The corresponding labels ('trainingLabels') are already formatted appropriately. After that, the dataset is divided into training and test sets using 'traintestsplit', assigning 20% of the data for testing purposes. The 'randomstate' parameter ensures reproducibility of results.

It is recommended to perform feature scaling for machine learning algorithms such as KNN that are affected by the magnitude of the data. This is achieved by standardizing the features using 'StandardScaler', which is fitted on the training and test sets.

The KNN model is initialized and trained using KNeighborsClassifier, with 'n-neighbors' set to 5. Training is conducted on the scaled training set. Post-training, the model is employed to predict outcomes on the test set. Model performance is evaluated using accuracy metrics, and a detailed classification report is generated.

Enhancement of the model is pursued through hyperparameter tuning using 'GridSearchCV'. A parameter grid is defined, comprising various configurations for 'n-neighbors', 'weights', and the distance metric. 'GridSearchCV' executes a cross-validated grid search across this parameter grid, identifying the optimal parameters.

Upon determination of the best parameters, the optimal model is trained. This model is identified as the most accurate during the grid search, and then it is used for predictions on the test set, and a classification report is produced.

Finally, the confusion matrix is calculated, and TP, TN, FP, and FN are extracted. These values compute sensitivity and specificity, offering insights into the model's efficacy in accurately identifying positive and negative classes.

This methodical approach ensures a thorough and effective implementation and assessment of the KNN algorithm, suitable for diverse classification tasks.

3.5 Experiments

3.5.1 MFCC and CNN

The new network draws inspiration from Rubin et al. [31]. There are similar functions with distinct properties in TensorFlow and PyTorch. For example, the padding procedure is called the same padding is not available in Pytorch. Then, torch.nn.ConstantPad2d is utilized before and after the convolutional layers to reach the same shape. In addition to this, the ReLU activation function is used after convolutional layers, and there are dropouts after fully connected layers. The authors use their custom-made loss function similar to Cross-Entropy Loss. Besides, they used L2-regularization. Batch-normalization can be used instead of L2 regularization. The new network has two batch normalization layers after the convolutional layers.

Inspired network failed in every attempt. Then the number of data is decreased to 800 samples. The network started to work, and %55-56 overall validation accuracy was achieved. The number of features is decreased to 12 and the number of data to 1000, then get %63 overall validation accuracy. The number of features is changed to 20 and the number of data to 1000; correspondingly, the network failed again. As a result, there is no linear relation between the number of features and records. The table 3.2 summarizes attempts. Then it is determined that the number of features is 12, and the number of data is 1000 samples.

Table 3.2: The Number of Record and Feature Attempts Summary

Number of Record	Number of Features	Validation Accuracy
3240	6	%50
800	6	%55
1000	12	%63
1000	20	%50

After that point, efforts were made to improve the results. The dropout values were initially set at 0.85565561, which had a negative impact on training performance by causing inputs to be excessively dropped and resulting in underfitting. To address this, the dropout values were reduced to 0.5 for both layers, and ReLU activation functions were added after the fully-connected layers. Additionally, changes were made to the optimizer; the SGD optimizer was used, and weight decay was initiated. With these adjustments, the network's training began, and the outcome was an overall validation accuracy of around 70%.

Hyper-parameter tuning was performed to optimize the model. Despite limited effects from modifying weight decay, the value was stabilized at 0.00008. Simultaneously, the learning rate was adjusted while maintaining constancy in other parameters. Eventually, a learning rate of 0.0001 was determined. The trials' outcomes are concisely summarized in Table 3.3.

Table 3.3: Learning Rate vs. Validation Accuracy Comparison

Learning Rate	Validation Accuracy
0.00015822	%72.33
0.0005	%67
0.001	%65
0.00005	%71
0.0001	%74

The highest achieved result is an approximate overall validation accuracy of 74%. The architecture of the optimal network during the forward pass is depicted in Figure 3.13.

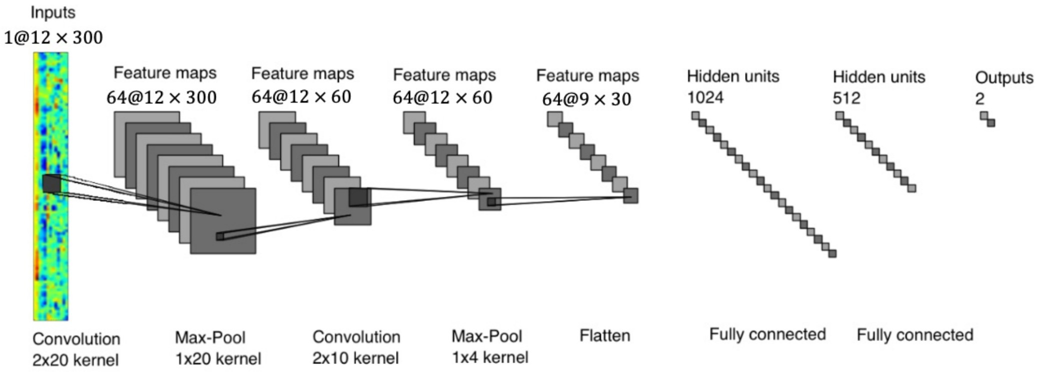


Figure 3.13: The New Convolutional Network Architecture adopted from [31]

The new hyper-parameters and the network parameters are summarized in Table 3.4 and Table 3.5.

Table 3.4: New Hyperparameter Values

Hyperparameter	Value
Learning Rate	0.0001
Weight Decay	0.00008
Dropout	0.5

Table 3.5: New Network Parameter Values

Network Parameter	Value
Regularization Type	Batch Normalization
Batch Size	256
Optimization	SGD Optimizer

The test dataset is not available then training and validation accuracy are used for comparison. In the best case, validation results before training are summarized in Table 3.6. After training, the results of the best case are represented in Table 3.7.

Table 3.6: Validation Results Before the Training

Normal Correct: 1	Normal Incorrect: 149
Normal Correct: 150	Abnormal Incorrect: 0
Sp = 1.0	Se = 0.00666666
Validation Accuracy Before Training: 50.33 %	

Table 3.7: Validation Results After the Training

Epoch: 101	
Normal Correct: 571	Normal Incorrect: 78
Normal Correct: 239	Abnormal Incorrect: 111
Sp = 0.6828571	Se = 0.8798151
Training Accuracy After Training: 78.13 %	
Normal Correct: 120	Normal Incorrect: 30
Normal Correct: 102	Abnormal Incorrect: 48
Sp = 0.68	Se = 0.8
Validation Accuracy After Training: 74.00 %	

The graph displaying the training and validation accuracy across epochs is given in Figure 3.14. In this Figure, the blue curve represents the training accuracy and the other one stands for validation accuracy.

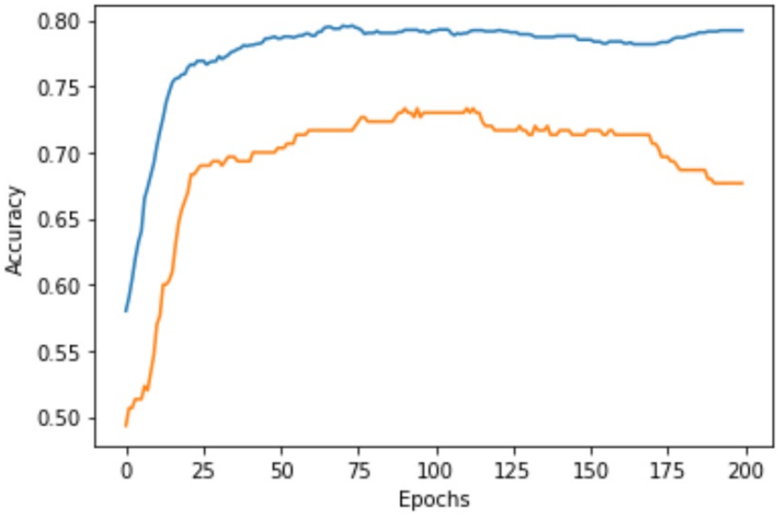


Figure 3.14: Epoch Number vs. Accuracy Graph

3.5.2 OpenL3 and SVM

In this part of the study, comprehensive experiments are conducted to examine the impact of data volume and kernel functions in the context of heart sound analysis.

In the analysis, the dataset comprises 985 samples divided into abnormal and normal categories. The abnormal category contains 448 samples, whereas the normal category contains 537 instances. This distribution is critical in understanding the dataset's composition and the challenges in effectively classifying the data.

A linear kernel SVM is employed for the classification task. The choice of a linear kernel is based on its suitability for high-dimensional data spaces, as it tends to perform well when there is a clear margin of separation between classes. Moreover, linear kernels are often preferred for their lower computational cost than other kernels.

The results of the SVM classification, reflecting the model's performance in distinguishing between normal and abnormal instances, are presented in Table 3.8. This table includes metrics such as accuracy, sensitivity, specificity, and other relevant statistical measures that provide insights into the effectiveness of the SVM model with a linear kernel in handling the given dataset.The best result of linear kernel SVM parameters are found as C = 0.01 and gamma = scale.

Table 3.8: Test Results After the Linear Kernel SVM Classification

Normal Correct: 71	Normal Incorrect: 16
Abnormal Correct:89	Abnormal Incorrect: 21
Sp = 0.77	Se = 0.85
Accuracy: 81.00 %	

Table 3.9 presents the experiments' results using an SVM with a polynomial kernel. The focus was the impact of varying the regularization parameter C ranging from 0.01 to 100 and different gamma values such as 'scale', 'auto', 1, and 10.

Table 3.9: Test Results After the Polynomial Kernel SVM Classification

Normal Correct: 54	Normal Incorrect: 35
Abnormal Correct:93	Abnormal Incorrect: 35
Sp = 0.60	Se = 0.86
Accuracy: 73.00 %	

The experiments reveal a notable trend in the performance metrics as C, and the best results are obtained when C = 100 and gamma = scale in SVM with a polynomial kernel. Although C increases and gamma changes, the specificity value does not increase enough through the experiments.

Table 3.10 displays the outcomes of experiments conducted using an SVM with an RBF kernel. These experiments These experiments desired to investigate the impacts of altering the regularization parameter C with values ranging from 0.01 to 100. Additionally, a variety of gamma settings were explored, including 'scale', 'auto', 1, and 10.

Table 3.10: Test Results After the RBF Kernel SVM Classification

Normal Correct: 64	Normal Incorrect: 15
Abnormal Correct:93	Abnormal Incorrect: 25
Sp = 0.72	Se = 0.86
Accuracy: 79.00 %	

The optimal results were achieved with C=100 and gamma set to 'auto'. Comparing the polynomial with an RBF kernel, the RBF kernel performs better. It shows a more balanced classification capability with fewer misclassifications of healthy cases, better identification of patients, and a higher overall accuracy.

3.5.3 OpenL3 and KNN

In this section of the research, the OpenL3 tool is employed for extracting features, and the k-Nearest Neighbors (KNN) algorithm is applied for classification. The dataset size remains consistent as 448 abnormal and 537 normal samples. Initially, Euclidean distance is the chosen metric. Additionally, various parameters are adjusted: the 'k' value is selected within a range of 3 to 15, and the weighting parameters are set to either 'uniform' or 'distance'. The primary focus during analysis is on optimizing accuracy.

The table referenced as 3.11 summarizes the best results obtained. The optimal configuration involved selecting three neighbors and setting the weighting parameter to 'uniform' in the scenario where the best results were achieved using Euclidean distance.

Table 3.11: Test Results After the Euclidian Distance KNN Classification

Normal Correct:79	Normal Incorrect:17
Abnormal Correct:89	Abnormal Incorrect:12
Sp = 0.87	Se = 0.84
Accuracy: 85.50 %	

The second attempt involves exploring the effects of different configurations in a method that uses the Manhattan distance approach. It systematically changes the number of neighbors from 3 up to 15. Alongside this, it alternates the strategy for weighing the influence of these neighbors between 'uniform' and 'distance'.

The table referenced as 3.12 delivers the best results. In Manhattan distance, the optimal configuration includes selecting three neighbors and setting the weighting parameter to 'uniform'. Compared to Euclidean, specificity and sensitivity are enhanced.

Table 3.12: Test Results After the Manhattan Distance KNN Classification

Normal Correct:80	Normal Incorrect: 16
Abnormal Correct:90	Abnormal Incorrect: 11
Sp = 0.88	Se = 0.85
Accuracy: 86.50 %	

The Minkowski distance metric is chosen in the third part. This metric is a more generalized form of the Euclidean and Manhattan distances. Like in the first two experiments, the "k" values range is considered, and the same approach to weighing parameters is employed.

The table referred to as 3.13 demonstrates the most effective outcomes. With the Minkowski distance metric, the superior configuration entails choosing p = 3, three neighbors and applying the ’uniform’ weighting strategy.

Table 3.13: Test Results After the Minkowski Distance KNN Classification

Normal Correct: 75	Normal Incorrect: 16
Abnormal Correct:93	Abnormal Incorrect: 13
Sp = 0.82	Se = 0.88
Accuracy: 85.30 %	

3.5.4 VGGish and SVM

This experiment investigates the application of the VGGish model coupled with a Support Vector Machine (SVM) for classifying heart sounds. This study is essential for enhancing our comprehension of the dataset’s characteristics and its classification challenges.

The dataset remains the same as in the previous study, consisting of 985 samples divided into two categories: 448 abnormal and 537 normal heart sounds. This distribution is essential for understanding the challenges in achieving accurate classification.

VGGish model was initially used for processing audio signals. VGGish converts raw audio data into high-dimensional feature representations, which could more effectively capture complex patterns than traditional methods.

These extracted features are then used to train an SVM classifier. The SVM is configured with various values for the regularization parameter ’C’ ranging from 0.001 to 100 and the kernel coefficient ’gamma’ including ’scale’, ’auto’, 1, and 10 while maintaining a linear kernel. This approach allows for a broader investigation of the model’s behavior under different parameter settings.

The performance of the SVM, using features extracted by VGGish, is documented in Table 3.14. The most effective parameter combination was found at ’C’ = 0.01 and

’gamma’ = scale. This finding highlights the effectiveness of combining advanced audio feature extraction with the computational efficiency of a linear kernel SVM.

Table 3.14: Test Results After the Linear Kernel SVM Classification

Normal Correct: 60	Normal Incorrect: 27
Abnormal Correct:81	Abnormal Incorrect: 29
Sp = 0.67	Se = 0.75
Accuracy: 71.00 %	

After that, the kernel type was changed into polynomial type. Initially, the regularization parameter ’C’ was set across a range from 0.01 to 100, and the kernel coefficient ’gamma’ at ’scale’, ’auto’, 1, and 10. The best results were initially observed with C = 0.01, gamma = 1. Consequently, it is decided to broaden the range of C values beyond the initial lower limit of 0.001 to explore whether further improvements in classification performance could be achieved.

The best result parameters did not change after the change. It is observed with C = 0.01 and gamma = 1 with the polynomial kernel. The performance of the SVM, utilizing the VGGish-extracted features and the polynomial kernel, is demonstrated in Table 3.15.

Table 3.15: Test Results After the Polynomial Kernel SVM Classification

Normal Correct:60	Normal Incorrect: 14
Abnormal Correct:94	Abnormal Incorrect: 29
Sp = 0.67	Se = 0.87
Accuracy: 77.00 %	

The last attempt is changing the kernel into the RBF type. Since the optimum results were obtained in the C = 100 value, the C value range was updated to 1000. Gamma parameters stay the same.

Interestingly, it was observed that increasing the C value beyond 100 did not significantly impact the results. The optimal performance was achieved with C = 100 and gamma = ’scale’, using the RBF kernel. The details of these results are presented in Table 3.16.

Table 3.16: Test Results After the RBF Kernel SVM Classification

Normal Correct: 64	Normal Incorrect: 15
Abnormal Correct:93	Abnormal Incorrect: 25
Sp = 0.72	Se = 0.86
Accuracy: 79.00 %	

3.5.5 VGGish and KNN

The VGGish model is utilized for feature extraction, and the k-Nearest Neighbors (KNN) algorithm is applied for classification in this part of the study. The dataset consists of 448 abnormal and 537 normal samples. The initial approach uses the Euclidean distance metric, with the ’k’ value varying from 3 to 15 and the weighting parameters set to either ’uniform’ or ’distance’. The primary objective is to optimize accuracy based on these parameters.

The results in Table 3.17 indicate the optimal configuration for Euclidean distance with ’n-neighbors’ set to 3 and the weighting parameter as ’uniform’.

Table 3.17: Test Results After the Euclidian Distance KNN Classification

Normal Correct:66	Normal Incorrect:17
Abnormal Correct:89	Abnormal Incorrect:25
Sp = 0.73	Se = 0.84
Accuracy: 78.50 %	

The Manhattan distance metric is examined further with the same range of 'k' values and weighting strategies. This approach investigates the impact of different distance calculations on classification accuracy with VGGish audio features.

As demonstrated in Table 3.18, the best performance using Manhattan distance was achieved with 'n-neighbors' set to 10 and the weighting parameter as 'distance'.

Table 3.18: Test Results After the Manhattan Distance KNN Classification

Normal Correct:61	Normal Incorrect: 17
Abnormal Correct:89	Abnormal Incorrect: 30
Sp = 0.67	Se = 0.84
Accuracy: 75.50 %	

Finally, the Minkowski distance metric, a generalized form of both Euclidean and Manhattan distances, is evaluated. The same parameters for 'k' values and weights are considered to assess the effectiveness of this metric.

Table 3.19 shows the results for Minkowski distance, where the most effective configuration was with 'n-neighbors' set to 3, the weighting parameter as 'uniform' and p value is 6. This method achieved 79.18% accuracy.

Table 3.19: Test Results After the Minkowski Distance KNN Classification

Normal Correct: 69	Normal Incorrect: 19
Abnormal Correct:87	Abnormal Incorrect: 22
Sp = 0.76	Se = 0.82
Accuracy: 79.20 %	

3.6 Discussion of the Results for Chapter 3

The Linear Kernel SVM, when applied to OpenL3 features, exhibits the highest overall accuracy at 81.00% in OpenL3 features. Its sensitivity, the ability to correctly identify true positives, is promising at 0.85. However, the specificity, which is the indicator of correctly identifying true negatives, stands at 0.77.

In contrast, the Polynomial Kernel SVM exhibits a lower overall accuracy of 73.00%, the least among the three kernels. Its specificity is notably lower at 0.60, suggesting a higher rate of false positives. However, its sensitivity is slightly superior at 0.86, implying a slightly better capability in identifying true positives.

Lastly, the RBF Kernel SVM strikes a balance with an accuracy of 79.00%, placing it between the linear and polynomial kernels. It has a better specificity result than the polynomial kernel.

Linear kernel's performance indicates a strong ability to identify true positives and a reasonably good rate at identifying true negatives, making it a solid choice for many applications. This suggests a robust performance in general classification tasks. The summary of the OpenL3 and SVM with different kernels is given in Table 3.20.

Table 3.20: Summary of OpenL3 and SVM Results with Different Kernels

Kernel Type	Accuracy (%)	Sp	Se
Linear Kernel	81.00	0.77	0.85
Polynomial Kernel	73.00	0.60	0.86
RBF Kernel	79.00	0.72	0.86

In addition to the accuracy approach, Figure 3.15 presents a Receiver Operating Characteristic(ROC) Curve. It compares the false positive and true positive rates.

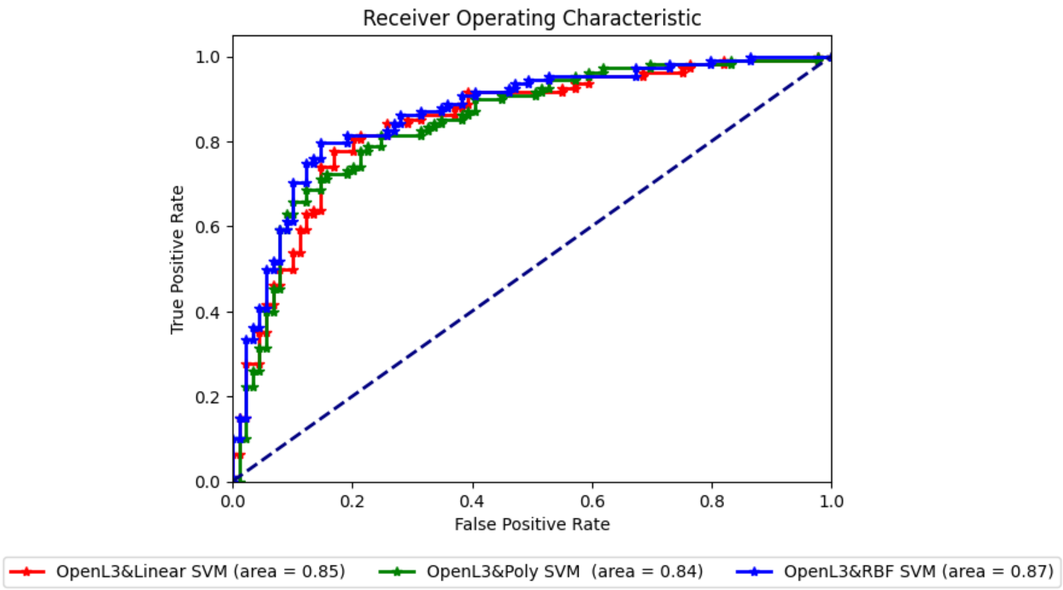


Figure 3.15: ROC Curve for OpenL3 and SVM with Different Kernels

The dashed line represents a random guess, where the true positive rate equals the false positive rate. A good classifier should be positioned far from this decision boundary, preferably towards the top-left corner. All three SVM models perform significantly better than a random guess, as indicated by their respective curves being closer to the top-left corner.

The area under the curve (AUC) is a measure of the classifier's ability to distinguish between the two classes. The AUC for each classifier is indicated in the legend. The RBF SVM has the highest AUC (0.87), suggesting the best overall performance among the three.

It is important to note that comparisons of accuracy values and ROC curve analyses may yield different results due to their distinct approaches to evaluation. The ROC curve evaluates the balance between the true positive and false positive rates, considering the specific needs of the application and the costs associated with misclassification.

The Euclidean Distance metric, when applied to OpenL3 features, shows a commendable performance with an accuracy of 85.50%. It demonstrates a strong capability in identifying true positives and true negatives, with values of 0.84 and 0.87. This indicates a reliable and balanced approach to classification tasks.

In comparison, the Manhattan Distance metric slightly outperforms the Euclidean Distance with an accuracy of 86.50%. This increase in accuracy is accompanied by improvements in both sensitivity and specificity, at 0.85 and 0.88. These results suggest a slightly better overall performance.

The Minkowski Distance, while having a slightly lower accuracy of 85.30%, shows an equivalent specificity to the Manhattan Distance at 0.88 but a slightly lower sensitivity of 0.82. This indicates that while the Minkowski Distance is as good as the Manhattan Distance in correctly identifying negative cases, it is slightly less effective in identifying positive cases.

The Manhattan Distance metric is the most effective in this context. It offers the highest accuracy and the best balance between sensitivity and specificity. The summary of the OpenL3 and KNN with different distance metrics is given in Table 3.21.

Table 3.21: Summary of OpenL3 and KNN Results with Different Distance Metrics

Distance Metric	Accuracy (%)	Sp	Se
Euclidean Distance	85.50	0.87	0.84
Manhattan Distance	86.50	0.88	0.85
Minkowski Distance	85.30	0.88	0.82

Figure 3.16 shows the ROC curve analysis that illustrates a superior AUC for the Euclidean KNN (0.93), indicating a robust ability to discriminate between classes. Despite having a lower accuracy than Manhattan, the high AUC value suggests that Euclidean KNN might outperform at certain threshold levels. The ROC results support the effectiveness of the Manhattan Distance metric but also highlight the potential of the Euclidean KNN in specific operational contexts.

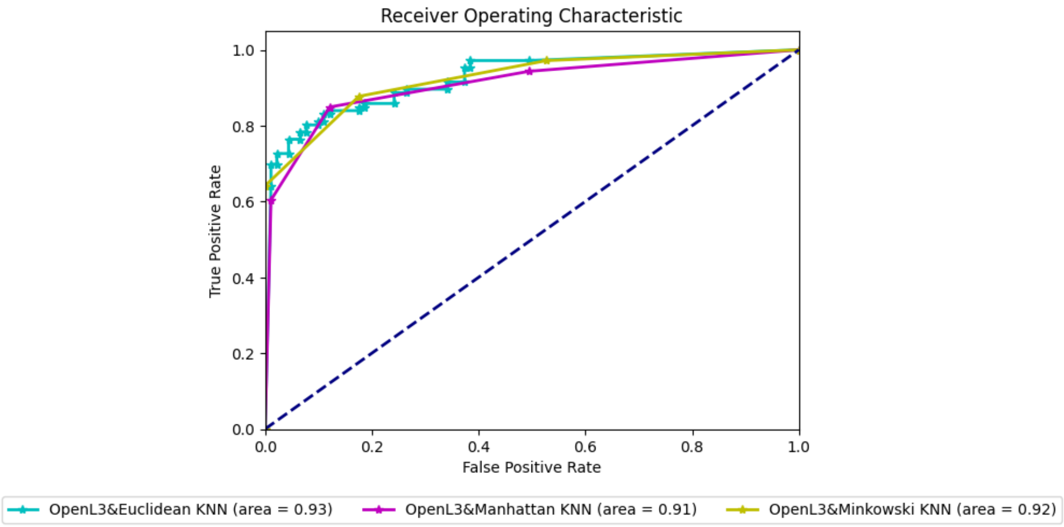


Figure 3.16: ROC Curve for OpenL3 and KNN with Different Distance Metrics

The Linear Kernel SVM shows moderate effectiveness with an accuracy of 71.00%. It has a specificity of 0.67 and a sensitivity of 0.75. This kernel type demonstrates balanced yet modest performance in classifying audio features.

On the other hand, the Polynomial Kernel SVM marks a notable improvement in accuracy, reaching 77.00%. Its specificity remains consistent with the Linear Kernel at 0.67, but there is a significant rise in sensitivity, reaching 0.87. While its ability to identify true negatives remains unchanged, the Polynomial Kernel performs better at correctly identifying true positives.

The RBF Kernel SVM shows the highest accuracy at 79.00%. It improves specificity to 0.72 and maintains a high sensitivity of 0.86. This balance suggests that the RBF Kernel is particularly effective in handling VGGish features and offers robust performance in identifying true positives and minimizing false positives.

The RBF Kernel SVM is the most effective choice for the SVM classification of VGGish features. The summary of the VGGish and SVM with different kernels is given in Table 3.22.

Table 3.22: Summary of VGGish and SVM Results with Different Kernels

Kernel Type	Accuracy (%)	Sp	Se
Linear Kernel	71.00	0.67	0.75
Polynomial Kernel	77.00	0.67	0.87
RBF Kernel	79.00	0.72	0.86

In Figure 3.17, the AUC values range from 0.79 for the Linear SVM to 0.84 for the RBF SVM, indicating that all models perform better than random chance. The RBF SVM, with the highest AUC of 0.84, is suggested to have the most robust discrimination ability among the three. The Polynomial SVM also shows a strong ability to discriminate, with an AUC of 0.83, closely following the RBF SVM. The Linear SVM, with the lowest AUC, could be valuable in specific operational contexts, especially considering factors like model complexity, computational efficiency, and interpretability.

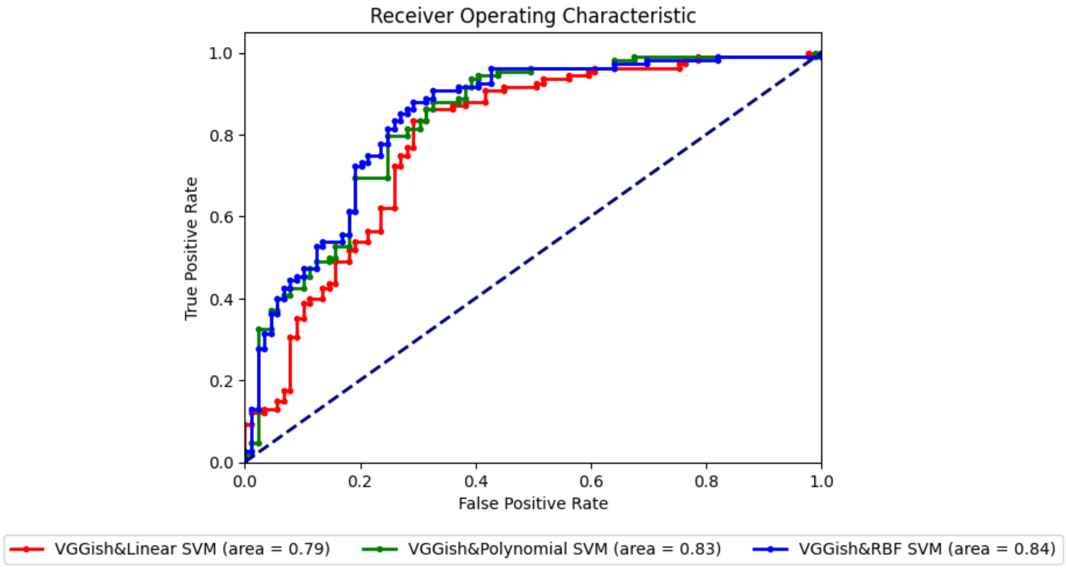


Figure 3.17: ROC Curve for VGGish and SVM with Different Kernels

While analyzing the VGGish and KNN results, the Minkowski Distance metric demonstrates a strong performance with an accuracy of 79.20%. This metric shows a solid balance between specificity at 0.76 and sensitivity at 0.82.

The Euclidean Distance metric shows a slightly lower overall accuracy at 78.50%. While maintaining the similar level of sensitivity as the Euclidean Distance at 0.84, it exhibits a lower specificity of 0.73.

The Manhattan Distance metric shows the lowest accuracy among the three at 75.50%. Despite this, it holds sensitivity the same with the Euclidean at 0.84 but falls behind in specificity with a score of 0.67.

Overall, Minkowski is the more effective choice for the KNN classification of VGGish features. The summary of the VGGish and KNN with different distance metrics is given in Table 3.23.

Table 3.23: Summary of VGGish and KNN Results with Different Distance Metrics

Distance Metric	Accuracy (%)	Sp	Se
Euclidean Distance	78.50	0.73	0.84
Manhattan Distance	75.50	0.67	0.84
Minkowski Distance	79.20	0.76	0.82

The ROC curve in Figure 3.18 illustrates the effectiveness of different KNN classifiers using VGGish features. The Euclidean KNN classifier demonstrates a strong discriminatory ability, with an AUC of 0.83. The Manhattan KNN performs slightly better than others with an AUC of 0.85, indicating a fine balance between true and false positive rates. The Minkowski KNN achieves the highest AUC of 0.86, suggesting it has the most robust discrimination capability of the three classifiers. These values, all above 0.8, demonstrate that each classifier performs significantly better than random chance, with the Minkowski KNN classifier being the most effective in this specific application for classifying VGGish features.

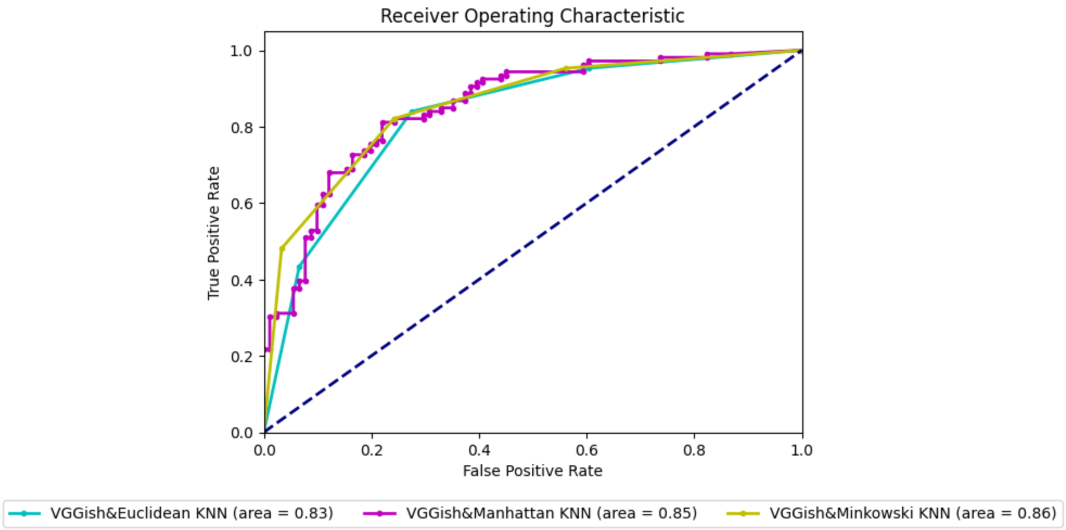


Figure 3.18: ROC Curve for VGGish and KNN with Different Distance Metrics

Table 3.24 compares five feature extraction method combinations: MFCC, OpenL3, VGGish, and classification algorithms, which are CNN, SVM, and KNN.

Table 3.24: Summary of the Best Results Across Different Methods

Method	Accuracy (%)	Sp	Se
MFCC & CNN	74.00	0.68	0.80
OpenL3 & SVM	81.00	0.77	0.85
OpenL3 & KNN	86.50	0.88	0.85
VGGish & SVM	79.00	0.72	0.86
VGGish & KNN	79.20	0.76	0.82

In MFCC & CNN, an accuracy of 74.00%, specificity of 0.68, and sensitivity of 0.80. While offering decent sensitivity, this combination shows lower overall accuracy and specificity than other methods.

The OpenL3 & KNN method outperforms others in accuracy, reaching 86.50%. The high specificity at 0.88 and sensitivity at 0.85 indicate that this combination is particularly effective at correctly classifying both negatives and positives. KNN with

OpenL3 features could catch the patterns more effectively, leading to these superior results.

In Figure 3.19, the ROC curves represent the performance of various classifiers using different feature extraction methods, as referenced from the results in the corresponding Table 3.24. The Manhattan KNN Classifer for the OpenL3 feature extraction method is very close to the upper left side and its AUC value is the highest, 0.91. On the other side, The MFCC and CNN has the closest line to the random guess curve and its AUC value is the lowest, 0.78.

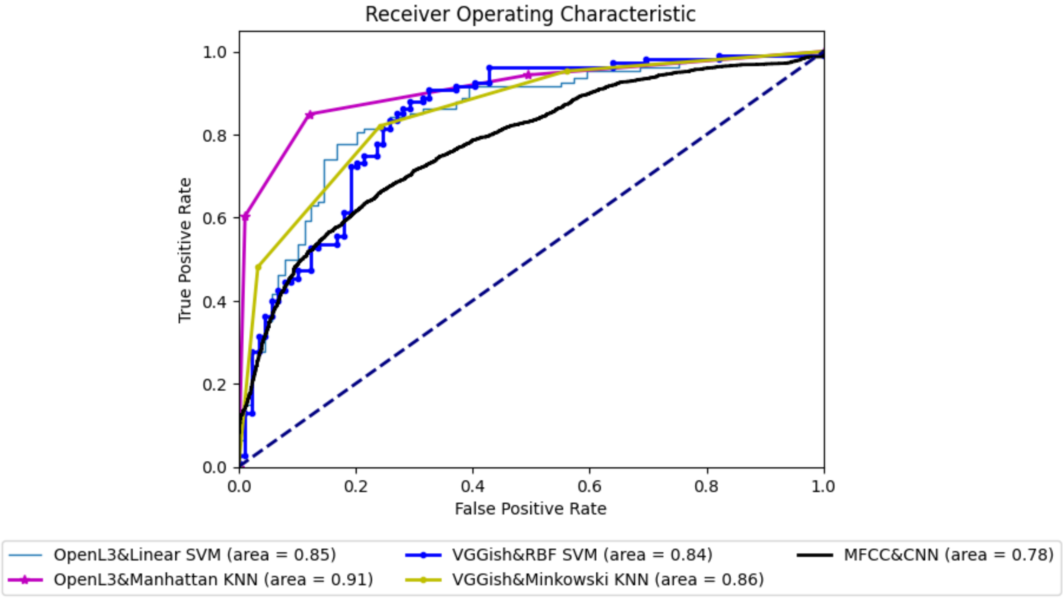


Figure 3.19: ROC Curve for the Best Results Across Different Methods

In conclusion, these findings underscore an important insight in audio classification: there is no need to build complex neural network structures to achieve good results. Instead, some fundamental and simple methods, as demonstrated by the OpenL3 and KNN approach, can accomplish high levels of classification accuracy and balance in performance metrics. It highlights the importance of method selection based on the specificities of the task rather than more complex solutions that may not always offer additional benefits.

CHAPTER 4

EXPLORING THE RELATION BETWEEN PHONOCARDIOGRAPHY AND VASCULAR SOUNDS

4.1 Vascular Sound Dataset

Tobin and Chang [21] showed a consistent relationship between the average pressure fluctuations experienced by the inner wall of a vessel and normalized distance independent of the Reynolds number.

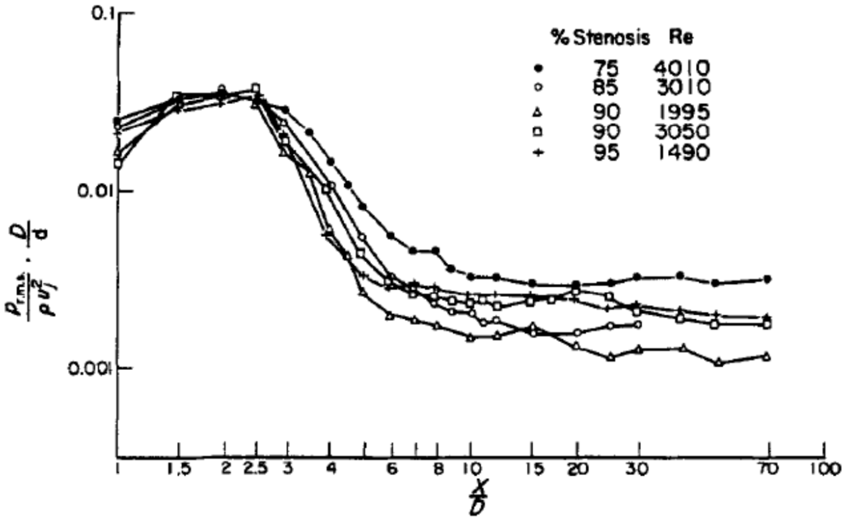


Figure 4.1: RMS Wall Pressure Fluctuations wrt x/D retrieved from [21]

Figure 4.1 is the root mean square (RMS) wall pressure fluctuations with x/D curves for different stenosis percentages. This figure demonstrates that there is a resemblance between the Reynolds number and stenosis level vs. the $\frac{p_{rms}}{\rho} \cdot \frac{D}{d}$ expression. Here, x represents the length measured from the end of the narrowed area in the flow direction. D is the tube diameter, while d stands for the constricted diameter. u_j is the average velocity at the narrowed area, ρ is the fluid density and p_{rms} shows the RMS pressure.

As Salman [57] mentioned, Yazıcıoğlu et al. [20] employed the equation 4.1 and Tobin and Chang's RMS pressure expression. They find the equation 4.2 by performing curve fitting in Matlab with the curves in Figure 4.1.

$$p(x) = 1.82 \cdot F_{n1}(x/D) \cdot \rho \cdot U^{3/2} \cdot \frac{D^{5/2}}{d^2} \left(\frac{1}{1 + 20(\frac{fd^2}{UD})^5.3} \right)^{1/2} \tag{4.1}$$

$$F_{n1}\left(\frac{x}{D}\right) = \frac{0.07057x + 0.3849}{x^2 - 23.22x + 167.9} \tag{4.2}$$

Salman [57] shared the source code that generates empirical results obtained from Tobin and Chang [21] study. The curve fit values 4.1 and the overall equation 4.3 are indicated as following:

Table 4.1: List of Pressure Distribution Equation's Parameters

Parameter	**Label**
$p_1 = 0.07057$	(1)
$p_2 = 0.3849$	(2)
$q_1 = -23.22$	(3)
$q_2 = 167.9$	(4)

$$p(x, f) = 1.82 \cdot \left(\frac{p_1 \cdot x + p_2}{x^2 + q_1 \cdot x + q_2} \right) \cdot 0.001 \cdot U^{1.5} \cdot \left(\frac{D^2}{d^2} \right) \cdot \left(1 + \left(20 \cdot \frac{f \cdot d^2}{U \cdot D} \right) \right)^{-0.5} \tag{4.3}$$

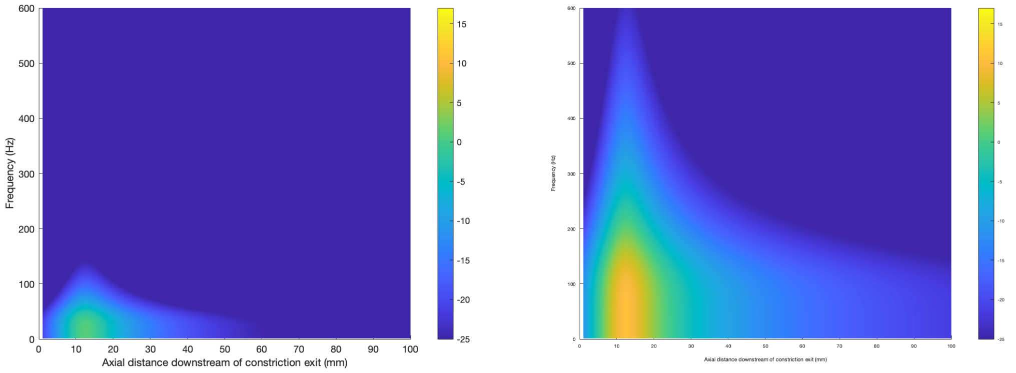
The equation referenced as 4.3 plays a crucial role in our analysis by providing an instantaneous pressure distribution. This distribution is not constant but varies depending on two key variables: the frequency of the measured signal and the 'x' value, which could represent the distance downstream from the exit of constriction.

To effectively demonstrate the pressure distribution within a vessel that has undergone 90% and 70% stenosis, Figure 4.2a and 4.2b present a detailed visual representation. These figures show that a high degree of narrowing impacts the vessel's pressure dynamics. The employed parameters are also outlined in Table 4.2.

Table 4.2: Summary of Pressure Distribution Generation Parameters

Parameter	**Description**	**Value**
x_1	Range (mm)	1 to 100
f_1	Frequency range (Hz)	1 to 600
D	Vessel diameter (mm)	6.4
U	Flow velocity (mm/s)	156

Figures 4.2a and 4.2b are contour plots that represent three dimensional data. The x-axis represents a spatial measurement in millimeters. It indicates the distance measured from the exit of a constriction. The y-axis represents frequency in hertz (Hz). The color gradient represents a pressure level in dB (ref 1 Pa) and the scale on the right side correlates the colors to numerical values. The colors range from dark blue to yellow, with blue representing lower values and yellow representing higher values.



(a) Instantaneous Pressure Distribution of % 70 Stenosis

(b) Instantaneous Pressure Distribution of % 90 Stenosis

Figure 4.2: Comparison between the Different Stenosis Severities

Figure 4.2b visualizes how a 70% vessel narrowing affects pressure. The pressure disturbances are more apparent closer to the constriction and tend to dissipate as the distance from the constriction increases.

Similarly, Figure 4.2a displays the pressure distribution for a more severe constriction of 90%. The color contours in this figure show more significant pressure variations due to the increased severity of the stenosis.

Instantaneous pressure is the pressure at a specific moment within the cardiovascular system. Each heartbeat generates a pulse and leads to a fluctuating pressure profile that varies temporally with each cardiac cycle. Systolic and diastolic phases characterize the pulsating flow. The ventricles contract, discharge blood into the arteries, and create a peak in pressure during the systole. The ventricles relax, and the pressure falls during diastole. The transition from instantaneous pressure to pulsating flow involves understanding how pressure varies during these cycles. Since the sine wave signal represents the cardiac cycle's behavior, the blood flow velocity is turned into a sine wave given in equation 4.4.

$$U = |U_{max} \cdot sin\left(2 \cdot \pi \cdot f_{heartbeat} \cdot t\right)| \tag{4.4}$$

In the equation 4.4, U defines the blood flow velocity as a function of time t, while U_{max} is the maximum velocity that the blood flow achieves during the cardiac cycle. The sine function is utilized to model the oscillatory nature of blood flow. $f_{heartbeat}$ is the frequency of the heartbeat, usually measured in beats per minute (bpm). For this equation, it is converted to Hz.

It is important to understand that sound is a mechanical wave resulting from particles' vibrations in a medium. These vibrations can be represented as a combination of different frequencies.

Fourier's theory declares that any complex waveform, including the sound of pulsating blood flow, can be decomposed into a sum of simple sinusoids of different frequencies, amplitudes, and phases. This is known as the Fourier transform. Conversely, the Inverse Fourier transform reconstructs the original signal from its sinusoidal components. The formulation of the inverse Fourier transforms is as follows:

$$g(t) = \int_{-\infty}^{\infty} G(f) e^{i2\pi ft} df \tag{4.5}$$

The model developed by Yazıcıoğlu et al. [20] and Salman [57], which is based on the study by Tobin and Chang [21], provides a fundamental basis for generating the sound associated with pulsating flow through a vessel.

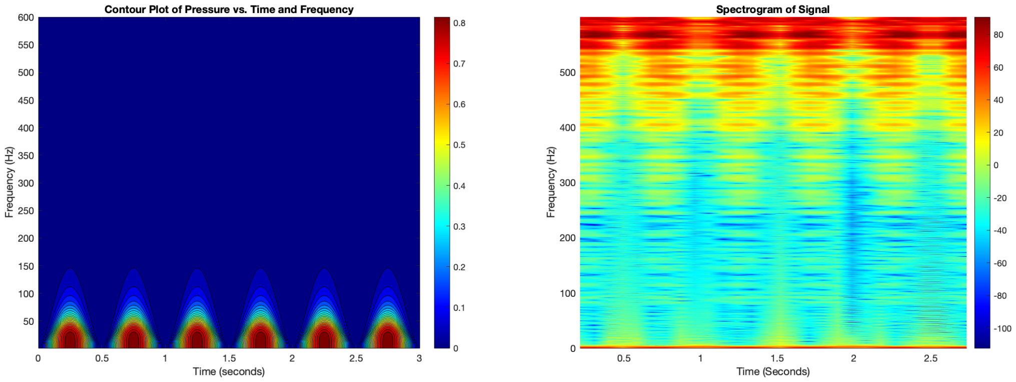
The sound waves result from the summation of various frequency components derived from pressure values. Since the model provides pressure values for a range of frequency values, the inverse Fourier transform could help generate the complex pulsating flow signal from the instantaneous pressure distribution. The total sound signal is generated by equation 4.6.

$$\text{signal} = \int_{1}^{600} p(t,f) \cdot e^{(i \cdot 2\pi f \cdot t + \phi)} df \tag{4.6}$$

In equation 4.6, p(t,f) originates from equation 4.3. As the maximum vibration occurs $x = 1.5 \cdot D$, the pressure expression's dependency is reduced to time and frequency. This equation resembles an inverse Fourier transform, which reconstructs a time-domain signal from its frequency-domain representation. The term ϕ represents a phase shift, crucial for accurately depicting oscillation.

In the figure 4.3a and 4.4a, the horizontal axis represents time that spans from 0 to 3 seconds. The vertical axis indicates frequency, which ranges from 0 to 600 Hz. The color intensity at any given point on the plot corresponds to the pressure magnitude and frequency at that time. Darker regions indicate lower pressure levels, whereas brighter colors signify higher ones.

The pattern of concentric and arch-like contours shows up periodically, suggesting a repetitive nature of the pressure variations over time. Red stands for higher intensity, and blue indicates lower intensity in the color bar. When comparing 70 % and 90 % stenosis levels, it is evident that 90% has higher pressure values.

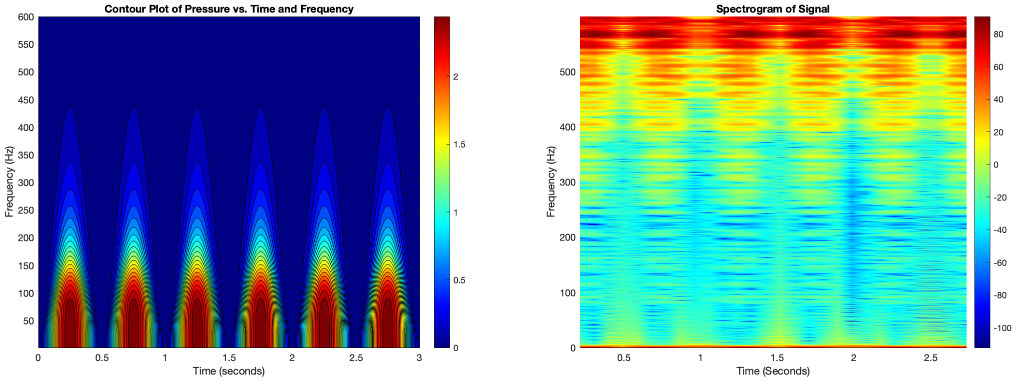


(a) Contour Plot of % 70 Stenosis Sound

(b) Spectrogram of % 70 Stenosis Sound

Figure 4.3: Comparative Visualization of Acoustic Characteristics for a 70% Stenosis Condition

The spectrograms 4.3b and 4.4b provide an understanding of a signal's frequency spectrum over time. They show the signal's intensity at various frequencies with bright colors. The scale on the right indicates the power of the signal in decibels. The horizontal axis range is between 0 and 2.5 seconds, and the frequency on the vertical axis extends up to 500 Hz.



(a) Contour Plot of % 90 Stenosis Sound

(b) Spectrogram of % 90 Stenosis Sound

Figure 4.4: Comparative Visualization of Acoustic Characteristics for a 90% Stenosis Condition

The red and yellow indicate higher intensity, whereas blue indicates lower intensity. The color intensity suggests the signal's energy distribution difference comparing 4.3b and 4.4b. In the level of 90%, more intense colors dominate the spectrum, indicating a signal with a consistently higher power across the measured frequencies and time intervals.

4.2 Relating Heart and Vascular Sounds

4.2.1 Evaluating with Pre-trained MFCC and CNN

In this section of the research, a pre-trained Convolutional Neural Network (CNN) is utilized. This CNN has been previously trained using the Physionet Heart Sound Database, which contains a comprehensive collection of heart sound recordings. The primary objective is to analyze the similarities between phonocardiography (heart sounds) and vascular sounds. Schematic illustration of the evaluation of vascular sounds using a pre-trained MFCC and CNN is given in Figure 4.5.

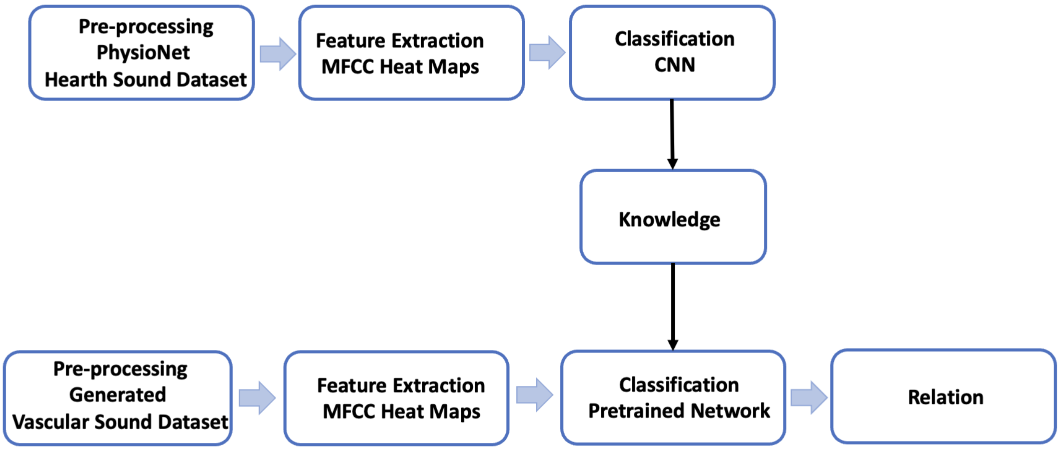


Figure 4.5: Schematic illustration of the evaluation of vascular sounds using a pre-trained MFCC and CNN

To achieve this, the vascular sound data is fed into the pre-trained network. In the process, just before the softmax layer of the network, probabilities indicating whether the sounds are normal or abnormal are extracted. This method provides insights into the comparative analysis of heart and vascular sounds.

The parameters of the vascular sound generation are given in Table 4.3.

Table 4.3: Summary of Sound Generation Parameters

Parameter	**Description**	**Value**
x_1	Range (mm)	1 to 100
f_1	Frequency range (Hz)	1 to 600
S	Stenosis Severity	30 to 95
D	Vessel diameter (mm)	6.4
U	Flow velocity (mm/s)	156

4.2.2 Evaluating with Feature Comparison

In this part of the study, the four databases of PhysioNet heart sounds and generated vascular sound dataset, are processed using the OpenL3 algorithm separately, the implementation of which was detailed in Chapter 2. This process extracts features, subsequently transforming the resulting matrices from 3-dimensional to 2-dimensional data.

The final phase involves normalizing the feature sets of both heart and vascular sounds. This normalization is achieved using the normalize function from the scikit-learn library, ensuring that the data sets are aligned on a standard scale. The normalized feature sets are then subjected to matrix multiplication to compute the cross-correlation matrix. This step is crucial for analyzing the relationships and similarities between the heart and vascular sound features.

The dimensions of the resulting cross-correlation matrix are printed, providing insight into the relational structure between the two datasets. Additionally, a count of instances where the cross-correlation value falls below a threshold (0.001 in this case) is computed and printed, offering a quantitative measure of similarity or dissimilarity between the heart and vascular sound features. The schematic illustration of the cross-correlation process using OpenL3 features of phonocardiography and Vascular sounds is given in Figure 4.6.

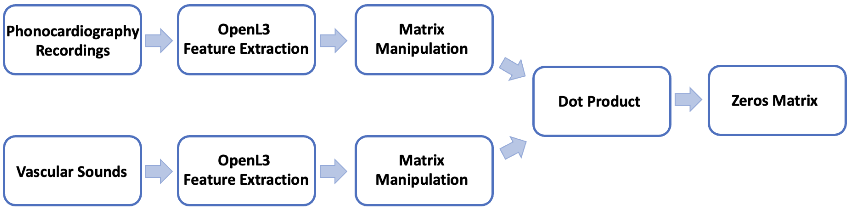

Phonocardiography Recordings
OpenL3 Feature Extraction
Matrix Manipulation
Dot Product
Zeros Matrix
Vascular Sounds
OpenL3 Feature Extraction
Matrix Manipulation

Figure 4.6: Schematic Illustration of the Cross-Correlation Process Using OpenL3 Features of Phonocardiography and Vascular Sounds

The same procedure was applied to the VGGish features, yielding results that were consistent with those obtained from the OpenL3 analysis. This consistency further verifies the findings derived from the OpenL3 approach.

4.3 Discussion of the Results for Chapter 4

The plot 4.7 presented is a scatter type overlaid with a fitted line. It is drawn to visualize the relationship between stenosis level and the probability of abnormality with the MFCC and pre-trained CNN combination.

The horizontal axis represents the stenosis level, quantified by a range from around 30 to 90. This range suggests a scaled measure of stenosis severity, with higher values indicating more significant narrowing.

The vertical axis shows the probability of abnormality, with values ranging from 0.4 to above 0.8. The probability scale is from 0 to 1, where 0 indicates no chance, and 1 indicates absolute certainty.

The blue dots represent individual data points that correlate a stenosis level with a probability of abnormality. These points are collected from the test of the pre-trained network with vascular sounds.

The orange line represents the trend within the data. The line looks like linear and there is an apparent relationship between the stenosis level and the probability of abnormality across the full range of data shown. Between 50% and 90% stenosis level, there is an increasing trend in probability but it is not a solid linear relationship.

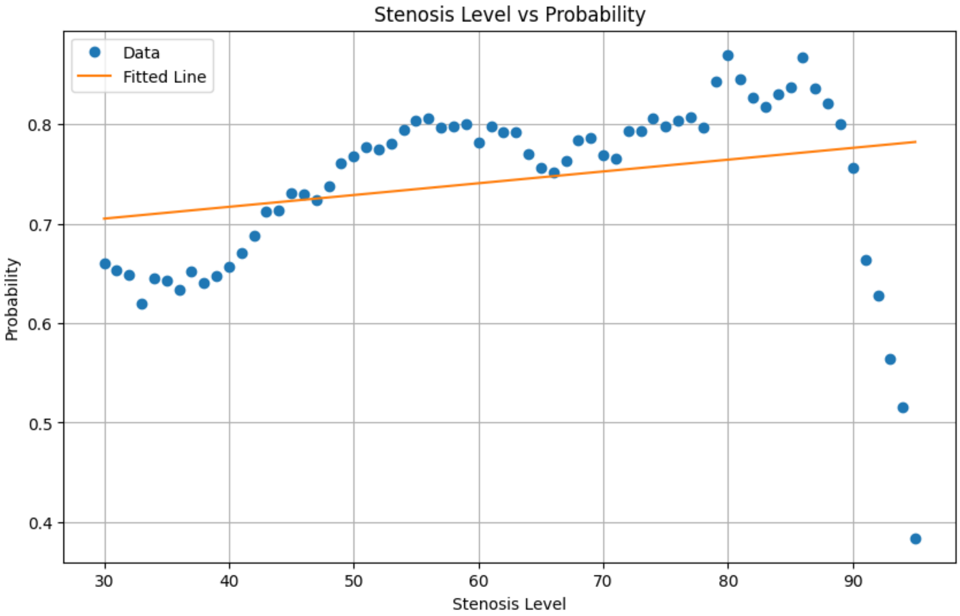


Figure 4.7: The Stenosis Level vs. Abnormal Probability Plot

The OpenL3 and VGGish analysis results that the computed cross-correlation matrix predominantly consists of zeros. This outcome suggests a lack of direct, linear relationship between the phonocardiography and vascular sound data set.

In conclusion, the analysis presented through the scatter plot with a fitted line offers a subtle understanding of the relationship between stenosis level and the probability of abnormality, as interpreted by the MFCC and a pre-trained CNN approach. The data, spanning a significant range of stenosis levels, reveal a pattern that is not strictly linear, indicating the complexity of vascular sound analysis. Furthermore, the results from OpenL3 and VGGish analyses, showing a predominantly zero-filled cross-correlation matrix, strengthen the absence of a direct, linear correlation between phonocardiography and vascular sound datasets. This study therefore contributes to a deeper understanding of vascular sound analysis, highlighting the need for sophisticated analytical techniques in accurately diagnosing and understanding vascular abnormalities.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This research has provided powerful insights into cardiovascular health diagnostics using phonocardiography (heart) and vascular sounds. Through comprehensive analyses employing different feature extraction methods (MFCC, OpenL3, VGGish) and classification algorithms (CNN, SVM, KNN), this study has demonstrated these combinations' varying levels of effectiveness in accurately classifying audio data.

The OpenL3 & KNN method emerged as the most effective, achieving the highest accuracy at 86.50%. Its superior performance in both specificity and sensitivity underscores its potential as a reliable method for audio classification tasks. In contrast, the MFCC & CNN combination showed lower overall accuracy and specificity, though it still maintained decent sensitivity.

In understanding the relationship between stenosis level and the probability of abnormality, the study revealed an apparent pattern between the 50-90% stenosis level. The scatter plot with a fitted line, particularly with MFCC and a pre-trained CNN combination, showed that the relationship is more complex than a straightforward linear correlation. In addition to that, the feature comparison method supports the findings with zero feature similarity. This finding is critical in advancing the understanding of vascular sound analysis and developing more effective diagnostic tools.

5.2 Future Work

In the pursuit of advancing the field of cardiovascular health diagnostics, various issues emerge for future research. First and foremost, exploring enhanced feature extraction techniques, including advanced deep learning approaches, holds the potential for even more accurate classifications. Broadening the scope of data sets to include diverse patient demographics and various stages of stenosis would also be invaluable in validating these findings.

Further, combining other variables such as age, gender, and patient history provides a more comprehensive understanding of phonocardiography recordings. The potential for implementing these classification methods in real-time diagnostic tools is an ongoing and promising research area. This greatly enhances the efficiency and effectiveness of medical diagnostics.

Comparative studies with existing diagnostic methods are crucial to strength classification techniques. The observed non-linear relationship between stenosis level and abnormality probability necessitates further investigation. More sophisticated mathematical modeling or machine learning techniques capable of capturing complex patterns in data could provide deeper insights into this phenomenon.

Expanding the scope of research, different combinations of feature extraction and classification methods would help evaluate the robustness. Combining the strengths of different techniques could yield more powerful and accurate systems.

Finally, researching ways to detect anomalies automatically in heart sounds through continuous monitoring could lead to early diagnosis and in-time intervention. This holds the potential to improve patient outcomes significantly. In a broader context, especially in regions with limited access to healthcare will be an important step considering the global health impact of these technologies.

In summary, phonocardiography in diagnostics has a bright future and opportunities. Integrating artificial intelligence, advanced data analysis techniques, and user-centered design principles holds the promise of revolutionizing how heart and vascular diseases are diagnosed and managed.

REFERENCES

[1] “World health organization,” 2023. Accessed on November 20, 2023.

[2] P. Libby, “The changing landscape of atherosclerosis,” 4 2021.

[3] P. Libby and G. K. Hansson, “From focal lipid storage to systemic inflammation: Jacc review topic of the week,” 9 2019.

[4] P. Libby, P. M. Ridker, and G. K. Hansson, “Progress and challenges in translating the biology of atherosclerosis,” *Nature*, vol. 473, no. 7347, pp. 317–325, 2011.

[5] J. Kakadiya, M. Jagdish, and L. Kakadiya, “Causes, symptoms, pathophysiology and diagnosis of atherosclerosis-a review address for correspondence,” 2009.

[6] G. S. Mintz, S. E. Nissen, W. D. Anderson, S. R. Bailey, R. Erbel, P. J. Fitzgerald, F. J. Pinto, K. Rosenfield, R. J. Siegel, E. M. Tuzcu, and P. G. Yock, “American college of cardiology clinical expert consensus document on standards for acquisition, measurement and reporting of intravascular ultrasound studies (ivus),” *European Journal of Echocardiography*, vol. 2, pp. 299–313, 2001.

[7] M. Y. Henein, S. Vancheri, G. Bajraktari, and F. Vancheri, “Coronary atherosclerosis imaging,” 2020.

[8] A. Giavarini, I. D. Kilic, A. R. Diéguez, G. Longo, I. Vandormael, N. Pareek, R. Kanyal, R. D. Silva, and C. D. Mario, “Intracoronary imaging,” *Heart*, vol. 103, pp. 708–725, 2017.

[9] T. Roleder, J. Jąkała, G. L. Kałuża, Ł. Partyka, K. Proniewska, E. Pociask, W. Zasada, W. Wojakowski, Z. Gąsior, and D. Dudek, “Review paper the basics of intravascular optical coherence tomography,” *Advances in Interventional*

Cardiology/Postępy w Kardiologii Interwencyjnej, vol. 11, no. 2, pp. 74–83, 2015.

[10] C. M. Gardner, H. Tan, E. L. Hull, J. B. Lisauskas, S. T. Sum, T. M. Meese, C. Jiang, S. P. Madden, J. D. Caplan, A. P. Burke, R. Virmani, J. Goldstein, and J. E. Muller, "Detection of lipid core coronary plaques in autopsy specimens with a novel catheter-based near-infrared spectroscopy system," 2008.

[11] G. S. Mintz and G. Guagliumi, "Intravascular imaging in coronary artery disease," 8 2017.

[12] S. Achenbach and P. Raggi, "Imaging of coronary atherosclerosis by computed tomography," *European heart journal*, vol. 31, no. 12, pp. 1442–1448, 2010.

[13] G. Cismaru, T. Serban, and A. Tirpe, "Ultrasound methods in the evaluation of atherosclerosis: From pathophysiology to clinic," *Biomedicines*, vol. 9, no. 4, p. 418, 2021.

[14] D. N. Ku, "Blood flow in arteries," *Annual review of fluid mechanics*, vol. 29, no. 1, pp. 399–434, 1997.

[15] W. Yongchareon and D. F. Young, "Initiation of turbulence in models of arterial stenoses," *Journal of Biomechanics*, vol. 12, no. 3, pp. 185–196, 1979.

[16] J. Fredberg, "Pseudo-sound generation at atherosclerotic constrictions in arteries," *Bulletin of mathematical biology*, vol. 36, pp. 143–155, 1974.

[17] G. W. Wijntjens, M. A. van Lavieren, T. P. van de Hoef, and J. J. Piek, "Physiological assessment of coronary stenosis: a view from the coronary microcirculation," 2015.

[18] R. S. Lees and C. F. Dewey Jr, "Phonoangiography: a new noninvasive diagnostic method for studying arterial disease," *Proceedings of the National Academy of Sciences*, vol. 67, no. 2, pp. 935–942, 1970.

[19] P. Bishop, "Evolution of the stethoscope," *Journal of the Royal Society of Medicine*, vol. 73, no. 6, pp. 448–456, 1980.

[20] Y. Yazicioglu, T. J. Royston, T. Spohnholtz, B. Martin, F. Loth, and H. S. Bassiouny, "Acoustic radiation from a fluid-filled, subsurface vascular tube with

internal turbulent flow due to a constriction," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1193–1209, 2005.

[21] R. J. Tobin and I.-D. Chang, "Wall pressure spectra scaling downstream of stenoses in steady tube flow," *Journal of Biomechanics*, vol. 9, no. 10, pp. 633–640, 1976.

[22] N. Freidoonimehr, R. Chin, A. Zander, and M. Arjomandi, "An experimental model for pressure drop evaluation in a stenosed coronary artery," *Physics of Fluids*, vol. 32, no. 2, 2020.

[23] H. E. Salman, "Non-invasive acoustic detection of vascular diseases from skin surface using computational techniques with fluid-structure interaction," 2018.

[24] K. Ozden, C. Sert, and Y. Yazicioglu, "Effect of stenosis shape on the sound emitted from a constricted blood vessel," *Medical & biological engineering & computing*, vol. 58, pp. 643–658, 2020.

[25] C. Sapsanis, N. Welsh, M. Pozin, G. Garreau, G. Tognetti, H. Bakhshaee, P. O. Pouliquen, R. Mitral, W. R. Thompson, and A. G. Andreou, "Stethovest: A simultaneous multichannel wearable system for cardiac acoustic mapping," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, IEEE, 2018.

[26] M. Klum, F. Leib, C. Oberschelp, D. Martens, A.-G. Pielmus, T. Tigges, T. Penzel, and R. Orglmeister, "Wearable multimodal stethoscope patch for wireless biosignal acquisition and long-term auscultation," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5781–5785, IEEE, 2019.

[27] J. Cui, Y. Li, Y. Yang, P. Shi, B. Wang, S. Wang, G. Zhang, and W. Zhang, "Design and optimization of mems heart sound sensor based on bionic structure," *Sensors and Actuators A: Physical*, vol. 333, p. 113188, 2022.

[28] P. Samanta, A. Pathak, K. Mandana, and G. Saha, "Classification of coronary artery diseased and normal subjects using multi-channel phonocardiogram signal," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 426–443, 2019.

[29] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Computers in biology and medicine*, vol. 111, p. 103346, 2019.

[30] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *2016 Computing in cardiology conference (CinC)*, pp. 609–612, IEEE, 2016.

[31] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Recognizing abnormal heart sounds using deep learning," *arXiv preprint arXiv:1707.04642*, 2017.

[32] A. Gupta, G. Tang, and S. Suresh, "Heartfit: An accurate platform for heart murmur diagnosis utilizing deep learning," *arXiv preprint arXiv:1907.11649*, 2019.

[33] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[34] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

[35] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[36] V. Despotovic, M. Ismael, M. Cornil, R. Mc Call, and G. Fagherazzi, "Detection of covid-19 from voice, cough and breathing patterns: Dataset and preliminary results," *Computers in Biology and Medicine*, vol. 138, p. 104944, 2021.

[37] V. Sheth, U. Tripathi, and A. Sharma, "A comparative analysis of machine learning algorithms for classification purpose," *Procedia Computer Science*, vol. 215, pp. 422–431, 2022.

[38] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling.," in *Ismir*, vol. 270, p. 11, Plymouth, MA, 2000.

[39] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856, IEEE, 2019.

[40] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[41] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, (New Orleans, LA), 2017.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[43] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[44] G. Castel-Branco, G. Falcao, and F. Perdigão, "Puremic: A new audio dataset for the classification of musical instruments based on convolutional neural networks," *Journal of Signal Processing Systems*, vol. 93, pp. 977–987, 2021.

[45] B. Mahmood, "What are artificial neural networks?," 2023. Accessed: 2023-12-01.

[46] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[47] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[49] "What is a support vector machine?."

[50] E. Fix and J. L. Hodges, "Nonparametric discrimination: consistency properties," *Randolph Field, Texas, Project*, pp. 21–49, 1951.

[51] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[52] IBM, "K-nearest neighbors (knn)."

[53] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson, *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological measurement*, vol. 37, no. 12, p. 2181, 2016.

[54] B. Karnath and W. Thornton, "Auscultation of the heart," *Hospital Physician*, vol. 38, no. 9, pp. 39–45, 2002.

[55] L. Narayana M and S. K. Kopparapu, "Choice of mel filter bank in computing mfcc of a resampled speech," *arXiv e-prints*, pp. arXiv–1410, 2014.

[56] Marl, "Marl/openl3: Openl3: Open-source deep audio and image embeddings."

[57] H. E. Salman, "Investigation of fluid structure interaction in cardiovascular system from diagnostic and pathological perspective," Master's thesis, Middle East Technical University, 2012.