AI-DRIVEN HEART HEALTH MONITORING SYSTEM USING A CUSTOM-BUILT DIGITAL STETHOSCOPE

by

Banneheka Mudiyanselage Lakindu Banula Sirimewan Banneheka

EC/2020/043

A Dissertation
Submitted to the
Department of Physics and Electronics
in partial fulfillment of the requirements for the
Degree of Bachelor of Honours in Electronics and Computer Science

Principal Supervisor : Prof. K.M.D.C. Jayathilaka [PhD (Colombo)]
Co – Supervisor : Senior Prof. S.R.D. Kalingamudali [PhD (Sheffield)]
Prof. A.L.A.K. Ranaweera [PhD (Kyung Hee)]

University of Kelaniya
Kelaniya, Sri Lanka
June, 2025

DECLARATION

I, B.M.L.B.S Banneheka, hereby declare that this thesis entitled AI-Driven Heart Health Monitoring System Using a Custom-built Digital Stethoscope is my original work and has not been submitted previously in whole or in part for the award of any degree. All sources used in this research have been properly cited and acknowledged.

Signature: _____ Date: _____

Supervisor Name	Signature	Date
Senior Prof. S.R.D. Kalingamudali		
Prof. A.L.A.K. Ranaweera		
Prof. K M D C Jayathilaka		

AI-Driven Health Monitoring System Using a Custom-Built Digital Stethoscope

B.M.L.B.S Banneheka
University of Kelaniya, 2025

Abstract

Cardiovascular disease remains one of the leading causes of death worldwide, especially in regions with low access to costly diagnostic testing. Outdated methods of assessment, which utilize stethoscopes, went largely unchanged for decades, with clinicians being heavily relied upon to report exactly what they heard from a patient, and heavily dependent on the expertise of the clinician. Because of this heavy reliance on expertise, early diagnosis of heart disorders can often suffer. What we present in this project is an ongoing work of a low-cost, AI-based digital stethoscope system that will capture the heart sounds and classify them as normal, abnormal, or artefact (non-cardiac noise) using machine learning features.

The device combines a commercially available USB-powered microphone and a 3D-printed stethoscope attachment. The heart sound recordings are processed with a Python-based pipeline that includes signal enhancement, noise reduction, and feature extraction. A machine learning model based on publicly available datasets (PhysioNet) had an overall 89% precision in multi-class classification with generally good class-wise accuracy for normal and abnormal categories.

The system exhibits great promise for low-resource settings in initial cardiac screening, but the current high-noise conditions weaken its performance. Future iterations will see real-time denoising algorithms and hardware improvements developed to provide better reliability and usability. The solution is reasonably inexpensive (approximately $30 per unit), scalable, and developed to fit rural healthcare workflows, making it an exciting prospect for telemedicine and first contact diagnosis.

Acknowledgments

I would like to express my deepest gratitude to everyone who has supported me throughout the course of this research project.

First and foremost, I am immensely thankful to my supervisors: the Dean of the Faculty of Science, Senior Professor Sudath R. D. Kalingamudali; Professor Charith Jayathilaka (Principal Supervisor, Department of Physics and Electronics, Faculty of Science, University of Kelaniya); and Professor Aruna Ranaweera (Department of Physics and Electronics, Faculty of Science, University of Kelaniya). Their invaluable insight, encouragement, and constructive guidance were instrumental in shaping this work.

My sincere thanks also go to Dr. Thoshini Kumarika (Department of Statistics and Computer Science, Faculty of Science, University of Kelaniya) and Dr. Hiruni Gunathilaka (Gampaha Wickramarachchi University of Indigenous Medicine) for their unwavering support, motivation, and critical feedback during the long months of this study. Their mentorship played a vital role in the successful completion of this research.

I gratefully acknowledge the assistance of the research team at the Electronic Device and Innovation Centre (EDIC), Faculty of Science, University of Kelaniya, Mr. Tharindu Gurusinghe, Mr. Pubudu Jayasekara, and Ms. Hiruni Gunewardana for their technical support and collaborative spirit.

I would also like to extend my heartfelt appreciation to my teammates, Pasindu Waidyarathan and Waruna Dissanayake, for their constructive ideas, dedication, and collaborative effort throughout this research journey.

Finally, I would like to thank all the volunteers who generously gave their time to participate in data collection, as well as everyone else who provided support and encouragement along the way. Your contributions have made this work possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

Introduction

1.0 Introduction

Cardiovascular disease (CVD) is the number one cause of global mortality, as 17.9 million deaths are caused by cardiovascular disease every year [14]. Instead, prevention and screening are parts of caring about morbidity and mortality through early detection. Direct reading of cardiac sounds (by using a stethoscope), which is known as auscultation, remains the main tool used by clinicians usually use to detect valvular lesions, murmurs, and other pathologies. The limitations to the diagnostic value of auscultation, however, are the long training period needed to develop the listening skills of an expert, as well as the degree to which there is subjectivity of sound. In low-resource environments and where skilled cardiologists and the use of sophisticated ECG technology are lacking, these constraints mostly result in delays or failures to make a diagnosis. Recent breakthroughs in machine-learning (ML)-driven acoustic classification have lately made an objective alternative in the form of digital signal acquisition. It is on this foundation that open-source ML and inexpensive micro-electronics are playing a significant role in creating cost-effective diagnostic devices despite the lack of publicly available heart-sound datasets. This paper, therefore, intends to design and prototype a low-cost portable digital stethoscope and software that can classify phonocardiograms as normal, abnormal, and artefactual, making it suitable to assist in cardiovascular screening in underserved settings.

1.1 Background of study

Cardiovascular diseases (CVDs) are the leading cause of death in the world, and it is of the essence that the clinical manifestation of these diseases be identified early to reduce their long-term effects. Conventional auscultation or listening to heart sounds using a stethoscope is easy to access and affordable in low-resource settings. Still, unreliability of listening to auscultatory tests, sizeable differences between observers, and the expertise of practitioners as a whole limit auscultatory diagnosis. Considering this, recently, the availability of low-cost digital signal processing and machine-learning environments has opened a potentially viable path towards augmenting traditional auscultation with intelligent systems that will reliably and reproducibly identify the presence of pathological murmurs. This paper discusses how to design and implement a low-cost and portable system that can analyse heart sounds. It is doing this by

using data-driven analysis methods like algorithms, models, and tools that are trained by collected data, such as machine learning.

1.2 Problem Statement

Although the global healthcare space continues to take giant steps, there is still obviously too little access to cardiology in most low and middle-income regions. Electrocardiography and echocardiography, the holy grail of diagnostic systems, have a high cost in terms of equipment and trained/experienced staff and constant electricity supply, which is unaffordable in even the more well-funded rural clinics or travelling clinics. Being hard pressed, clinicians use acoustic stethoscopes, the effectiveness of which requires the experienced ear to detect the slightest, situational heart sounds and, therefore, the results are likely to be subject to significant inter-observer variance and increase the risk of a mild early murmur being missed.

The issue is addressed to some extent by commercial digital stethoscopes that amplify voice and filter out peripheral noise but are still too expensive to be purchased by most resource-restricted departments. Affordable devices are even rare and those available rarely have machine-learning capabilities with the ability to adjust to the diverse acoustic environments and provide unbiased pathology callings in real time. Failure to close these gaps spells out a great need to have an AI-powered auscultation platform that makes high-fidelity cardiac signals with effective ambient noise suppression and is capable of identifying pathology with confidence on the fly, a platform which may be cardiac auscultation lifesaving to so many patients.

1.3 Objectives

A current inquiry aims to design, test, and implement an exceedingly low-cost computerised stethoscope and an integrated artificial intelligence structure to perform a primary screening of cardiograms in low-income circumstances.

- Create a stethoscope prototype powered by an audio jack and made of parts that can be adjusted off-the-shelf, along with a 3D-printed housing whose realistic use demonstrated the ability to record audible cardiac sounds consistently within the frequency range of 20 to 400 Hz.

- Annotated phonocardiograms (about 2,200 recordings) were also compiled, similar to the existing public datasets, but enhanced by including additional clinically validated samples (collected by a newly created digital stethoscope).
- A pipeline process involving signal processing was created in order to remove noise, segment and transform the signals into time-frequency features (MFCCs) that can be used in machine-learning processes.
- A combination of a convolutional neural network (CNN) to learn spatial relationships and recurrent units to learn temporal relationships, an RCNN deep network, was trained to classify recordings labelled as normal, murmur or artefact. The model was tested in terms of accuracy, precision, recall and F1-score with its performance results measured against past research and field-generated results.
- Developing a cloud-based (with AWS-EC2), web platform (using Next.js and Flask) that is capable of recording heart sound, noise reduction, executing real-time AI-based inference, and delivering real-time diagnostic feedback with the help of the trained model.

1.4 Scope and Limitations

The proposed work is a system for recognising heart sounds in three categories: normal, abnormal (pathological) and artefact (non-cardiac noise). The solution is a set with a low-cost digital stethoscope that is attached through an audio jack to a laptop, a tablet or a mobile phone. The current version cannot support wireless integration, though it is technically possible.

Signal analysis, noise reduction and signal identification are made via a cloud-based web interface, with the results of inference giving a preliminary notifications style, that is given by the signal analysis application framework and the result via a cloud-based web interface (e.g. due to abnormal heart sound, referral of a specialist is advised). The model is trained with some combination of publicly available databases (predominantly PhysioNet), and approximately 100 newly recorded normal heart sounds in healthy adults.

Testing was restricted to situations with a minimal amount of noise. When ambient noise levels are high, performance can be reduced. Data and inference in embedded devices and pediatrics (age < 18) do not fall in this category. Although developed as a research and demonstration system, the system has a simple interface so it would be suitable in early screening and testing.

1.5 Significance of the study

The World Health Organisation (WHO) puts the death cases of cardiovascular diseases at about 17.9 million a year, which is about 31% of the total deaths in the whole world [14]. Three-quarters of these deaths happen in developing nations, in which the complex diagnostic devices like echocardiography and ECG are relatively few. Conventional auscultation using an acoustic stethoscope is the most common screening tool available and research has shown that the sensitivity in detecting pathological murmurs is less than 50% in non-specialists [15].

The presented investigation is a portable digital stethoscope that can analyze and classify heartbeats into three categories automatically and cloud-based, with a cost of around 30 USD. This system will fill a severe diagnostic backlog of, (i) consistent and standardised in the initial diagnosis of abnormal heart sounds, (ii) elimination of redundant referrals and expenses, including cost of unnecessary physician appointments, and (iii) bringing fundamental heart screening to remote or rural low populated areas. Moreover, the selected and carefully labeled dataset that is produced in the framework of this project is made publicly available to the community to make further contributions to the development of machine-learning-based cardiac diagnosis.

1.6 Literature Review

Using automated analysis of heart sounds is crucial in an early diagnosis of cardiovascular illness considering that evaluation of heart sounds is subjective since it is done by listening by means of stethoscope[1], [2]. Traditionally, this is an area that uses conventional machine learning models, namely Artificial Neural Networks (ANN) and Support Vector Machines (SVM), following a lot of signal processing and feature extraction[3], [4], [5]. The recent emergence of deep learning (DL) subfield, especially CNNs and RNNs, has revolutionised the field to a significant level by allowing automatic extraction of complex features directly off of phonocardiograms (PCGs) [3], [6], [7], [8]. New research done using these architectures has claimed an accuracy of over 90% even with accuracies of over 99.8 % in some of their models [3], [5], [6], [9]. Moreover, the introduction of massive user-generated data, especially the PhysioNet/CinC Challenges, has driven this study and created possibilities of automated diagnostic tools and smart wearables [1], [3], [5], [6], [7], [9], [10].

1.6.1 Review of Existing Studies

Machine-Learning (ML) and Deep-Learning (DL) capabilities of analyzing heart sounds without the use of stethoscopes represent a non-invasive instead of subjective auscultation of cardiovascular illnesses [7], [9]. The main aim of this field is to design smart systems with the ability to identify diseases fast and accurately [9].

Convolutional Neural Networks (CNN) are an attractive structure of heart sound classification, which provides a strong performance on data that has structure or patterns in space, such as images or spectrograms (which are visual representations of sound) [7]. They have been combined in a hybrid framework, e.g. Convolutional Recurrent Neural Networks (CRNNs), where local features are coded by a 2D-CNN and longer-range temporal relations are extracted via a long short-term memory layer (LSTM). Values achieved on the database PhysioNet/CinC Challenge 2016 provide very good results of accuracy 98% and 86.8 % respectively, as the binary and multiclass ones [1], [2], [3]. CNNs have also been utilised by the surveys to determine several conditions like murmurs, valvular heart disease (VHD), congenital heart disease (CHD), heart failure (HF), coronary artery disease (CAD) and rheumatic heart disease (RHD) [6].

DL shows better recognition results than the classical methods of signal processing and is more economically viable and intuitive to the user than the standard medical image processing [6], [7]. The DL models are able to read small trends in PCG signals that a human ear is incapable of distinguishing [6]. Though it has been shown that combinations of Support Vector Machines (SVM) and Linear Predictive Coding (LPC) coefficients and Modified Cuckoo Search (MCS) exhibit good classification accuracy (over 93 % with 12 classes of heart sounds) [9], but in most cases, DL models can get equally good results with fewer computation times [3], [11]. However, apparently, DL architectures have their own drawbacks, such as limited interpretability and being prone to overfitting, particularly to smaller-sized datasets [1], [2], [6], [7].

The great number of works related to the usage of spectrograms, Mel-spectrograms or Mel-frequency cepstral coefficients (MFCCs) area or Mel-spectrograms take place in recent studies [1], [6]. Other discrete wavelet transform (DWT) characteristics are often exploited together with MFCCs [1]. The two combined have been demonstrated to greatly improve classification accuracy and reach all-time highs of 97.9 % when the Yaseen Khan 2018 dataset is processed on an SVM classifier [9], [12]. Remarkably, it is possible to learn end-to-end on the raw PCG signals, although some CNN structures are less suitable in applying raw audio as the input [2].

An alternative approach that does not involve the segmentation process can make a significant difference and save a lot of time at the cost of maintaining almost the same level of performance, for an example by calculating the wavelet entropy or the spectral amplitude [1], [11].

An adequate literature search of existing sets of heart sounds has revealed that the most notable ones are PhysioNet/CinC Challenge 2016, PASCAL, Yaseen Khan 2018, and PhysioNet Challenge 2022 [1], [6], [7], [13]. Both of these data sets have their own limitations, but together they restrict the use of the information:

The data related to PCG signals are prone to background noise, and in order to remove the noise, is it necessary to filter noise while in preprocessing stage [1], [4], [13]. Imbalance of classes: A number of PCG datasets are biased such that most of the recordings are normal, a factor which would make the models inaccurate [1], [2]. The biggest datasets are still relatively small in comparison with the capacity of current deep-learning models and therefore promote overfitting and reduce generalization [1], [2], [5], [6], [7]. Some of the datasets mark the sound as either normal or abnormal, do not allow making precise diagnoses of diseases [2], [5]. To ameliorate the limitations, researchers consider approaches including data augmentation, weighted loss functions and cross-validation methods to improve model performance and generalization [1], [2], [3], [6], [7].

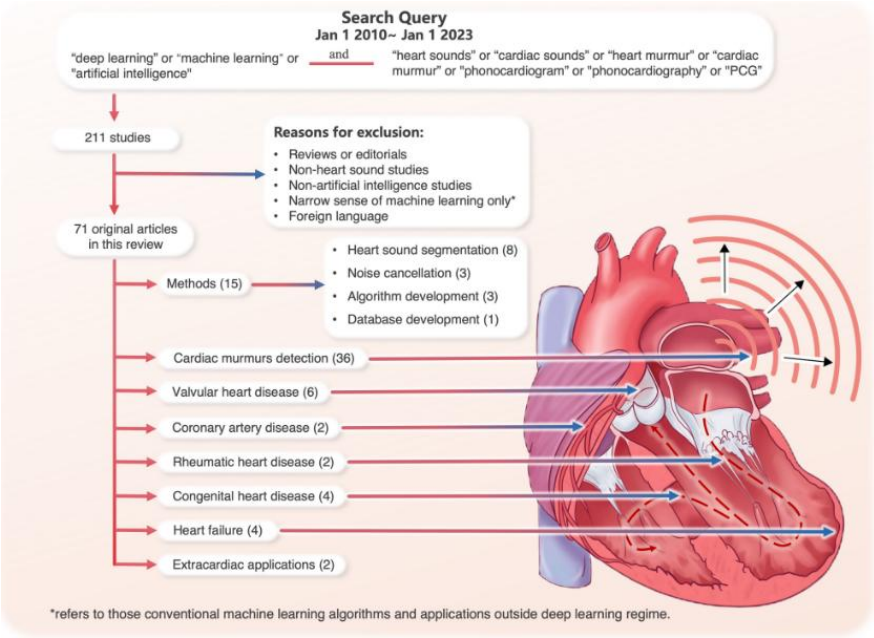

Figure 1.6.1 – Diagram demonstrating the literature-selection process and applications of deep learning in heart-sound analysis (Zhao et al., 2023)

1.6.2 Theoretical and Technical Foundations

Different heart sound classification procedures have been based on conventional methodologies, which depend on deep learning (DL) frameworks. The fundamental model is the Convolutional Neural Networks (CNNs), and it uses convolutional layers and receptive fields to learn the local features of spatial data representations like spectrograms, Mel-spectrograms, or scalograms. End-to-end architectures have the ability to accept raw phonocardiography (PCG) data, but some centered CNN designs prove to be better at dealing with pre-processed, image-based signals [2], [7].

Supervised learning has become a kind of reigning paradigm as models are trained to learn from labeled data. To address the most common class imbalance seen in medical related areas, the loss functions often use class weights to train the model [2]. The performance models are measured with accuracy, sensitivity, specificity, and F1-score and mean accuracy which is also useful because it combines sensitivity with specificity as well as measures performance [2], especially in imbalanced tasks, The popular approaches to practice that are aimed to validate the model and to avoid overfitting is to apply validation strategies namely, k-fold cross-validation, especially in the cases where data are limited or imbalanced [1], [2], [3].

The struggles with small and unequal data are partly overcome by data augmentation methods, such as data resampling, or what is known as synthetic data generation [1]. The drop out layers and Global Average Pooling (GAP) layers are also crucial as they simplify the model, decrease the chances of overfitting and make the model robust since they substitute the fully connected layers [2].

1.6.3 Research Gap Identification

Although significant milestones in processing and machine-learning methods of phonocardiography (PCG) signal analysis have been achieved, the challenges faced by scientists in establishing a range of critical research gaps that would facilitate the conversion of AI-facilitated cardiac screening to low-cost, non-invasive diagnostic devices remain eminent. Moreover, data scarcity, especially on well-labeled and diverse pathological samples, is a consistent problem that restricts the training of sound deep-learning models and tends to rule over-fitting and result in minimal transfer to unused clinical data.[1], [2], [3], [5], [6], [7], [9], [10].

The state of advancements in artificial intelligence in cardiac screening now faces a series of major restrictions that limit practical implementations. First, the use of complex deep learning models permeates the latency limitation, thus slowing down the actual time functioning and timely diagnostic reporting, which are critical components of point-of-care screening [3], [4], [9], [11]. Also, there seems to be a certain architectural conservatism, to the extent that a significant percentage of the literature still refers to traditional network topologies, which may be missing better hybrid or use-specific networks to better suit the non-stationary nature of PCG signals that are highly complex [1], [2], [3], [6], [7], [9], [10], [13]. At the same time, considerable obstacles that can be described as translational still exist and are based on the background noise and biased datasets, where these factors negatively affect the effectiveness, feasibility, and usability of intelligent stethoscopes or mobile health apps in practice [1], [2], [4], [5], [6], [7], [9]. The filling of these gaps will become a necessity in order to unlock the full potential of AI in cardiac screening.

1.6.4 Contribution of the Current Study

Our paper gives a solution to low-cost, AI-facilitated cardiac screening by improving the currently using AI systems. First, it proves out a low-cost digital stethoscope, using some commonly available components, and a custom 3D-printed enclosure, shows reliable acquisition of heart sounds over the range 20-400 Hz in field deployment conditions. Second, the research will compile and publish a clinician-annotated corpus of phonocardiograms involving 2,200 records, hence expanding the pathology variation. Third, an adaptive signal-processing pipeline with a modular interface to signal-processing modules is presented to perform denoising and time-frequency conversion (to MFCC), and which guarantees model-agnostic feature quality. Fourth, a hybrid convolutional-recurrent deep network design optimized to non-stationary acoustic dynamics is detailed by the study, attaining state-of-the-art accuracy, precision, recall and F1 on both internal and benchmark data. Fifth, a web-based app interacting with a cloud provides an inference performance in a couple of seconds and intuitive feedback visualizing real-time feasibility. Lastly, a scalability analysis and a roadmap to the assimilation of telehealth are identified, and this brings a translational gap between academic prototyping to the actual use of primary-care diagnostics.

CHAPTER II

Methodology

2.0 Introduction to Methodology

To illustrate that the triaging capability of heart sounds based on a very low-cost digital auscultation can achieve reliable results when combined with a lightweight deep-learning tool, this research follows a completely secondary-data engineering-oriented approach to its findings. Applied experimental activities rest upon a 2200-phonocardiogram (PCG) dataset freely available and mainly directly from the PhysioNet/CinC 2016 challenge and two other repositories of Kaggle, resulting in a balanced content, including 850 normal, 850 abnormal, and 650 artefact clips representing appr. 900 individual subjects. We will use only existing patients who are not recruited specifically. Hence, the level of ethical risk is minimal. Around 100 datasets are collected for testing purposes and for finding the usability of the newly developed digital stethoscope. Thus, the ethical risk is low.

A digital stethoscope was made, and it is created by an omnidirectional miniature condenser microphone physically attached to a regular stethoscope head, broadcasts its audio signal to a Web browser at 22.05 kHz 16-bit mono using the Web-Audio API to mimic the field experience. This is the combination of 5s WAV training a 128x128 single image under log-MFCC L2 masking with SpecAugment (time masking and frequency masking) and fed into a 2D CNN with residual blocks that has a total of 3.3M parameters. Supervised training is conducted through 50 epochs with Adam ($lr = 1x10^{-4}$), L2 regularisation, dropout and early stopping. The model has managed to reach a classification accuracy of 90% in general.

The Keras network is trained and containerised on a Flask microservice with the capability of CORS-origin running on an AWS EC2 instance. A React front-end allows a user to record the heart sounds directly or upload the already existing WAV files, which can then give output of the diagnosis as a popup alert after sending and processing the signal wave in flask API on AWS instance, the output will show normal, abnormal or artefact, with the confidence of the prediction. Such mixed pipeline raises the grounds of the subsequent evaluation of test accuracy, openness, and fiscal feasibility.

2.1 Research Design

This is a grounded design, practical-improvement probe that integrates secondary data analysis with rapid prototyping on hardware in order to close the diagnostic divide in low-resource cardiology. There are four consecutive layers of the study. For dataset creation, a consistent corpus of 2,200 publicly accessible phonocardiograms that is proportionate in terms of being (850 normal, 850 abnormal, and 650 artefact clips) was created to achieve the parity of classes and reduce the sampling bias. Then, for signal-processing experimentation, all five-second recordings were resampled to 22.05 kHz and down-converted into a 128 x 128 log-MFCC spectrogram, so that temporal-spectral structure was maintained, but direct compatibility with two-dimensional convolutional kernels was easily achieved. After that, a lightweight 3.3 million parameter 2D ResNet-like model was trained in a supervised setting using Adam optimisation, early stopping and dropout regularisation. The cross-validation of this computational experiment used an 80-20 stratified split in order to measure generalizability. Finally, the trained net was containerised into a Flask microservice and linked to a React front-end which allows the clinician to record or upload WAV files and be informed of the triage results in less than 15s.

The design is clear in that it is intervention oriented as it seeks to determine the ability of a budget stethoscope pipeline (architected on a browser) to demonstrate clinically meaningful accuracy and response time. Open-source tooling, pre-processing steps which can be easily audited, and the use of verifiably published datasets allow the maximum level of reproducibility to be achieved.

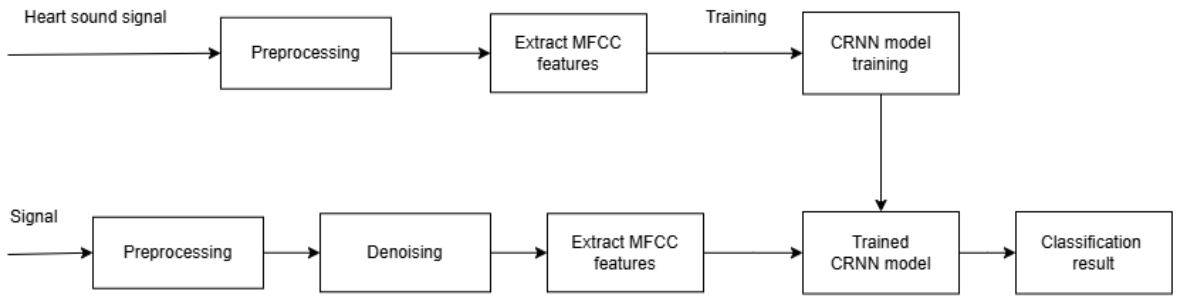

Figure 1.1.1 - The flowchart of the heart sound classification algorithm with convolutional residual neural network

2.2 Hardware Framework

As the given investigation shows, even sound cardiac structures that are essentially intelligible in clinics can be recorded by a probe assembly, the major elements of which, even though not as expensive as commercial electronic stethoscopes, are considerably cheaper. The probe is a miniature electret condenser capsule (size of 4.5 mm 2.2 mm; omnidirection polar pattern, nominal sensitivity of –36 dBV ± 3 dB (≈ 15.8 mV Pa^{-1})) and a frequency band of 50 Hz 20 kHz which includes the frequency range of 20 hertz and 800 hertz where most of the first and second-heart sounds, murmurs and common artefacts occur. Since the electret transducer is naturally designed to have an internal Field-Effect-Transistor (FET) buffer, the discrete amplifier circuit has become unnecessary. The capsule is hard-connected to a CTIA-standard TRRS plug, thus drawing the 2.0–2.7V plug-in bias present on laptop or smartphone audio codecs, the host pre-amps provide about 40–50 dB of gain.

The flexible tubing of the stethoscope was modified (by cutting the extruded length 45 mm above the chest-piece) in order to house the sensor (microphone), which effectively leaves a remaining front section extending through the diaphragm, maintaining an air column. The portion of the tube that had been cut out was substituted by a tailor-made two-component ABS adapter that was printed by an FDM printer in 15% infill (*Figure 2.2.4*). The adapter is made up of (i) a 20 mm x 39 cap module that fits over the remaining residual tubing, (ii) a 15 mm x 36 insert module which press-fits the electret capsule flush against the cap module using a threaded interface for stability. The reason why the two cavity openings of the tube, the cap module and the inset module in this setting should be precisely aligned is to ensure there is a coaxial relation between the acoustic port and the tube part, after which coupling gel or mechanical gaskets are not necessary.

Figure 2.2.1 – fully assembled probe – chest piece, microphone with 3d printed parts and cable

Figure 2.2.2 - parts of the probe – chest piece, microphone with 3d printed parts and cable

It is based on digital capture and thus no barb or continuation was made on proximal tubing, so standard binaural listening components were omitted. The strain relief and plug-and-play package enable the 30 cm male-to-female 3.5 mm TRRS extension cable connector to easily connect to audio jacks with smartphone or laptop applications. It was inspired by open-source, GliaX/Stethogram projets [16] [17].

Figure 2.2.3 - Assembled probe in use: smartphone → TRRS cable → microphone module → chest-piece resting on a volunteer.

Figure 2.2.4 - Rendered insert module (left), threaded cap module (right) and microphone (middle)

A brief bill of materials shows the cost of a capsule unit of USD 1.80, cost of chest-piece of USD 25.00, cost of a printed adapter of USD 1.00 and cabling plus sundries cost of USD 1.50. When added together, these elements can provide the total cost of a unit to be about USD .00. This probe (resulting) therefore meets the two design challenges of low-cost and high-quality phonocardiography sound acquisition.

2.2.1 Electret-Condenser Capsule

Although Section 2.2 shows that an off the shelf electret condenser capsule delivers sufficient acoustic bandwidth, it is worth unpacking to see the on-board electronics that make this possible. Each capsule has a depletion-mode junction gate field-effect transistor (JFET) configured as a source follower buffer. The diaphragm-electret pair behaves as a variable capacitor that produces nano-ampere charge variations; without buffering the capsule impedance exceeds 1 GΩ, rendering the signal unusable over practical cable runs. The internal JFET presents a gate impedance in the tera-ohm range, translating the extremely small charge modulation into a low-impedance (<2 kΩ) voltage swing at the drain, thereby protecting the audio band (50 Hz–20 kHz) information from capacitive loading or RF pick-up.

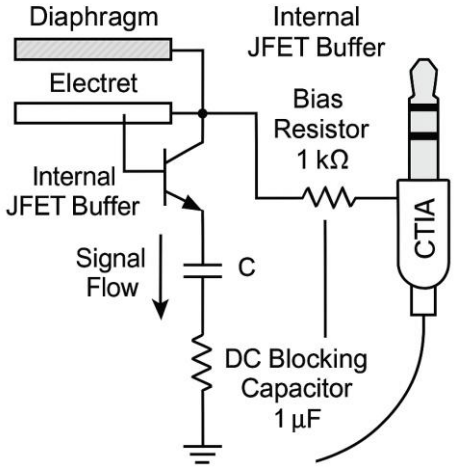

Figure 2.2.5 - Electret-Condenser Capsule: Internal Buffer & Bias Network

Figure 2.2.6 - Commercial Electret-Condenser Capsule

Biasing topology

The canonical single-supply bias network, adapting it to the smartphone and laptop codec systems that use the CTIA TRRS pin-out, is displayed in Figure 2.2.5. The capsule is driven by the 2.0 2.7 V plug-in power rail provided by the codec microphone bias pin and a load resistor (680 Ω – 2.2 kΩ; 1 kΩ typ.). This resistor serves two purposes: it determines the quiescent value of the drain current (0.2-0.5 mA, depending on the capsule) and converts the drain-current variations into a voltage that rides on the bias rail. This Voltage is AC-coupled to the pre-amp input by a DC-blocking capacitor (1 µF typ.) that eliminates the bias offset, but leaves the sub-50 Hz content to maintain the integrity of the S1 and S2 envelopes. The node impedance is then about 1kΩ at 20 Hz, it is well within the input range of commodity audio interfaces.

Implications for the prototype

- No discrete pre-amplifier required: Since the codec already implements a low noise 40 - 50 dB gain, an op-amp section can be eliminated in the probe.
- Cable-induced artefacts minimised: The low-impedance source follower tolerates the additional 30 cm shielded extension without appreciable high-frequency loss.

2.3 Inclusion & Exclusion Criteria

That was ensured by the rigid eligibility criteria to ensure the data at many levels reflect the requirements of real practice, adult bedside auscultation in conditions of limited resources, and to exclude the label noise and technical artefacts as much as possible.

Inclusion criteria

- To allow easy analysis and provide homogeneity of the data, we limited the data set to the recordings of subjects aged 18 years and above. All audio clips that correlated to confirmed adult subjects were included, and paediatric subjects were not used due to high heterogeneity.
- Any waveform was specifically and uniquely classified to one of three categories: (1) normal; (2) abnormal (any pathological murmur or extrasystole); or (3) artefact (non-cardiac noise). The waveforms that could not be classified as ambiguous or having multiple labels were dropped.
- During the current work, audio files of duration greater than 5.0 s were chosen so that they allowed for at least one cardiac cycle to be captured when padding and segmentation procedures were done at bradycardic rates.

Exclusion criteria

- Paediatric (< 18 y) or un-age-annotated patients.
- Recording clips of less than 5s,
- files containing header errors or severe clipping.

The usage of such filters delivered 2,200 PCG recordings (850 normal, 850 abnormal and 650 artefact) of 794 different adults. Since all the data is open-access, and all anonymisation was complete, no further institutional review was necessary. The following standards provide a balancing factor between the parity of the classes, the technical and ethical quality of the models, and an executable base on how to proceed regarding the modelling pipeline seen in the following sections.

2.4 Dataset & Sample Size

It was trained and tested on a custom-validated series of 2,200 phonocardiogram (PCG) recordings gathered solely in open-access repositories. The core of the dataset was the PhysioNet/Computing-in-Cardiology 2016 Challenge archive, which provides a body of 764 individually identified adult patients (P_0001 to P_0764). In order to bring variety to the classes, the Kaggle Phonocardiogram Heartbeat set was then ingested. All the rest of the 30 novel files, which were missing identifiers of patients, were considered new adults, which increased the effective number of subjects. One of the Kaggle sets consists of hospital ambient noise that provided non-cardiac waveforms, which were re-labelled as artefact, since these are environmental recordings and never provide subjects. To settle the unique in-house collection of 100 healthy-volunteer recordings was made with heart sounds of volunteers from the Faculty of Science, University of Kelaniya.

Having performed de-duplication, age filtering (18 y and over), and exclusion of clips of less than five seconds, the class balance is roughly 850 normal, 850 abnormal, and 650 artefact recordings, lessening optimisation bias but maintaining a sufficient sample size per class. This corpus was stratified at the patient level into corpus partitions by training (80%), validation (20%). Considering that the number of parameters in this network (3.3 M) was modest, from a earlier power analysis suggested that sample size of over 600 samples per class would provide greater than 95 % power of detecting a 5-point difference in accuracy at $\alpha = 0.05$. This means that the current sample size qualifies sufficiently on the basis of statistical rigour and external comparability with earlier studies done on PCG-classification.

2.5 Signal Pre-processing Pipeline

Each recording was equally treated within the software to discriminate the performance variation that can be dependent only on the model architecture. The raw WAV files were imported with librosa.load(sr = 22,050 Hz), enabling them to be converted to a floating-point array with the range normalised to -1 … 1. No pre-processing was carried out, for example, no explicit denoising, no band-pass filtering, or gain control. Rather, the convolutional layers of the network take charge of the filtering in the frequency domain, hence significantly reducing phase distortion.

Both waveforms in question are zero-padded or centre-truncated to a length of 10s, which is long enough to contain at least two cardiac cycles even in cases of a bradycardic pace (<= 40

bpm) without being excessively long in order to keep the GPU memory demands within a reasonable range. The resulting 500-sample length "vector" is then divided into 92.9ms frames using an FFT window length of 2048 and hop length of 512, providing around 430 of these frames per clip. The Spectral features are calculated in each frame with a 64-channel Mel filterbank (n mels = 64) sequenced by discrete cosine transform, with the first 64 cepstral coefficients preserved. The square of the log magnitude yields a 64 x 430 log-MFCC matrix that captures the information of the spectral envelope and the temporality dynamics in a simplified manner.

The MFCC matrix is then bilinearly down-sampled to a resolution of 128 by 128 pixels to fit the 1: 1 aspect ratio convention of 2-D convolutional neural-network kernels and is later written to an in-memory tensor that has the shape (1, 128, 128). SpecAugment on-the-fly is then applied, per mini-batches, a pair of temporally random masks (with maximum width of 10 frames) and a pair of frequency masks (with maximum width of 8 Mel bins) are overlaid. No additive white noise or gain jittering, or time-stretch processing is done, and the augmentation process is directed at making the signal more robust to spectral dropouts not amplitude variations.

This simpler pipeline (raw capture, temporal standardisation, MFCC projection, geometric resizing and lightweight spectral masking) produces input tensors that retain more diagnostically important content but with a much smaller volumetric data demand compared to the original audio data, hence leading to faster model convergence and inference.

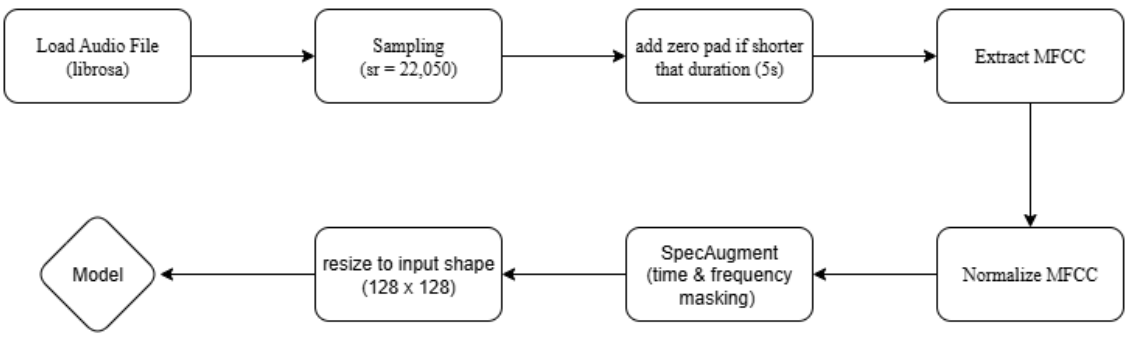

Figure 3.5.1 - Data Loading and Preprocessing Flow - Training and Prediction Processes

2.6 Model Development and Evaluation Strategy

This section describes an extended system for building a lightweight deep-learning classifier and describes the evaluation procedure that has been developed to assess its generalising performance.

2.6.1 Network Topology

A small 2-D Residual Convolutional Neural Network (RCNN) with a specific architecture, which is intentionally designed to work with 128×128 log-MFCC features input (Section 2.5), is used. The stem contains only one layer of 3 X 3 convolutions with 32 filters and then Batch Normalisation, ReLU activation layer, and Max-Pooling 2 X 2. This way, even global spectral-temporal edges are retained, yet spatial resolution is lowered.

The feature extractor uses three residual blocks with the depth of filters increasing progressively to 64, 128 and 256. Every block is the combination of two 3 x 3 Conv2D layers with further connecting Batch Normalisation and ReLU. Their Conv2D layers bring the sum of their output to the shortcut excitation, hence ensuring internal covariate stability. The dropout layers ($p = 0.20, 0.3, 0.4$) are added after each block so that features cannot co-adapt to each other. Taken together, the model has about 3.3 million trainable parameters.

Global Average Pooling is used to provide an intermediate representation of global features, and it removes position sensitivity and decreases parameters compared to full connected flattening. The resultant characterisation is grouped in an input to a 256-size dense layer with ReLU activation and Dropout ($p = 0.50$). This is followed by a 3-neuron Softmax layer, which gives mutually exclusive outputs of normal, abnormal and artefact concepts. All convolutional kernels are subjected to an L2 kernel regularizer ($\lambda = 10^{-4}$), which provides additional regularisation to the Dropout layer, enhancing out-of-sample performance.

Table 2.6.1 - Three-block residual CNN architecture for 128 × 128 log-MFCC inputs (≈ 3.3 M parameters)

Stage	Layers added	Cumulative count
Input	Input	1
Stem	Conv (1) → BN (1) → ReLU (1) → MaxPool (1)	5
Residual block 1 (64 filters)	Conv, BN, ReLU, Dropout, Conv, BN, Add, ReLU	13
Residual block 2 (128 filters)	Conv, BN, ReLU, Dropout, Conv, BN, shortcut Conv + BN, Add, ReLU	23
Residual block 3 (256 filters)	Conv, BN, ReLU, Dropout, Conv, BN, shortcut Conv + BN, Add, ReLU	33
Head	GlobalAvgPool, Dense (256), Dropout, Dense (soft-max)	37

2.6.2 Training Regime

Trained optimisation was done on a multi-core CPU host in TensorFlow/Keras 2.x with supervision. At the patient level, the dataset (Section 2.4) was stratified once and randomly divided by about 80 % and 20 % into two categories of training and validation, respectively, to avoid inter-fold dependence of beats that belong to a particular patient. The loss was categorical cross-entropy, which was minimised with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and an initial learning rate of 1×10^{-4}.

In addition, it only underwent the training up to 50 epochs, but an Early-Stopping callback (patience = 10, restore_best_weights=True) was applied to stop the optimisation process once the validation loss had stagnated, thereby avoiding over-fitting. A ReduceLROnPlateau callback reduced the learning rate by half after every three epochs when the validation loss had not dropped, and this allowed convergence around the optimum with no need to manually tune the schedule

2.6.3 Evaluation Protocol

Instead of k-fold cross-validation, only one validation set (20 %) was kept and represented an approximate situation during deployment, where the model would be exposed to other unseen patients only once. The framework saved the curves in the accuracy/ loss curve of training and validation accuracy/loss to permit divergent visual inspection of being exploited. Upon convergence, the frozen checkpoint with the best-performing checkpoint was kept and subsequently employed in inference testing.

Evaluation measures were extracted based on Scikit-learns, classification-report, reporting macro-averaged precision, recall, as well as F1-score, as well as per-class measures bringing relevance to the cost between murmur sensitivity and artefact specificity. Moreover, the true-/false-positives/negatives as absolutes were also documented in the confusion matrix. Since the dataset was balanced by class (around 850:850:650), the macro measurements were unbiased and sufficient to summarise it; weighted measurements and area under the receiver operating characteristic curve were referred to as redundant in this initial study.

2.6.4 Regularisation and Robustness Motivation

There were three orthogonal strategies to generalise:

1. L2 weight decay deterred convolutional filters of large scales with a tendency to over-specialisation.
2. In the dense head, and dropout after each residual block prevented co-adaptation and was a form of implicit model averaging.
3. SpecAugment synthetically masks random time and frequency regions, encouraging the network to rely on distributed contextual cues rather than narrow spectral peaks, a critical property when recordings originate from different stethoscope positions and varying conditions.

2.7 Deployment Architecture

Following the optimisation of the convolutional-neural-network (CNN) classifier of phonocardiogram (PCG) data, the real-time inference of it was engineered at a loosely-coupled, cloud-native deployment stack via a browser-based interface. The front-end uses Next.js and runs on Vercel so that client-side rendering has a minimum latency on the user experience. Then, the back-end uses Flask and an Amazon EC2 node so that computational flexibility is achieved.

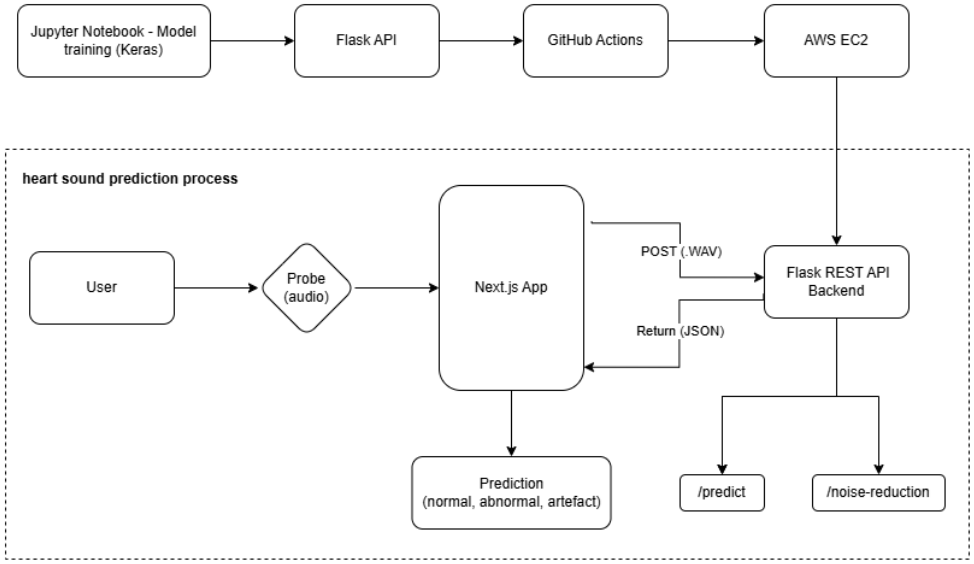

Figure 2.7.1. *End-to-end deployment pipeline: Jupyter-trained model → EC2-hosted Flask API (/noise-reduction, /predict) → Next.js client for real-time WAV upload and diagnosis process.*

Model hosting and API layer

The trained Keras model is encapsulated in a Flask application that exposes two REST endpoints: POST /noise-reduction, which applies spectral subtraction to remove ambient noise from the audio signals, and POST /predict, which returns the categorical distribution: normal, abnormal, artefact, enabling automated classification of heart sound recordings.

The service is containerised with Docker and continuous integration and deployment (CI/CD) pipelines defined in GitHub Actions. It is responsible for building the image, pushing version-tagged artefacts to EC2, and triggering zero-downtime redeployments on the EC2 instance.

On the EC2 host, Nginx reverse proxy terminates TLS via Let's Encrypt, performs request routing, and forwards traffic to a Gunicorn process pool, enabling horizontal scaling through an Auto Scaling Group if demand warrants.

Front-end delivery

The user interface is implemented in Next.js (v15) and deployed on Vercel's edge network. Client-side components leverage the Web Audio API to record 5s (or 10s, 15s, 30s) WAV clips or accept file uploads. Upon submission, the clip is first dispatched to /noise-reduction; the denoised waveform is then streamed to /predict. A modal dialogue immediately visualises the returned confidence scores and predicted label, while an asynchronous toast logs the inference latency to promote transparency.

Figure 2.7.2 – front-end application interface

End-to-end request flow (Figure 2.7.1)

1. A User can record or upload a heart-sound clip on the Vercel-hosted UI.
2. The clip is transmitted over HTTPS to a Dockerized environment on EC2.
3. Then it forwards to the request to the Flask container's /noise-reduction, which returns a cleaned WAV.
4. The cleaned clip is piped to /predict; the RCNN infers and serialises the probability vector.
5. The UI receives the JSON payload and renders the prediction and accuracy.

2.8 Summery

The following is my description of the entire process that was followed to make a low-cost acoustic sensor a clinically relevant diagnostic test. I started a pragmatic research design that was both low in cost and time. I expand on this scheme by presenting the details of a stethoscope probe quantity to connect to a stethoscope and reproduce its sound as a digital stream at 22.05 kHz without an additional amplifier. The $29 probe consists of an electret capsule tightly fitted into a press-fitted ABS adapter produced with 3D printing. A discriminating list of inclusion and exclusion criteria left me with 2,200 curated audio clips of some 794 distinct patients, which were longer than 5s, and the patients involved were adults with unambiguous labels.

Then, I create a balanced dataset that contains about 850 normal, 850 abnormal and 650 artefact clips that are all divided patient-wise into 80/20 folds to avoid data leakage (Section 2.4). In order to compress the gigantic audio data, I use a reproducible pipeline whose first element is the temporal normalisation, followed by Mel spectral projection and finally bilinear resampling together with SpecAugment data masking, reducing the data size while yet not losing the diagnostic information (Section 2.5).

On machine learning, trained with dropout and L2 weight decay a 3-block residual CNN (3.3 M parameters), (Section 2.6). Train the net with feature-optimiser Adam, early-stopping and dynamic learning-rate decline, train and also report stable convergence on common CPUs. The evaluation is done on a held-out validation set, and their macro-precision, recall, and F1 scores are obtained along with a confusion matrix, indicating the confusion against classes.

Finally, Section 2.7 outlined a cloud-native deployment architecture: a Dockerised Flask micro-service on AWS EC2 exposes /noise-reduction and /predict endpoints behind a Nginx TLS proxy, while a Vercel-hosted Next.js interface records or uploads WAV clips and presents results. CI/CD pipelines in GitHub Actions ensure reproducible, zero-downtime updates.

Overall, this chapter presents an open, replicable pipeline, pathway, and endpoint, which in this case includes the ability to capture sensors and deliver inference over the web, with cost and performance challenges. The second chapter determines the diagnostic value of the model, contrasts it with previous studies, as well as examines the latency, robustness, and economic viability of the model within the restrictions of real-world factors.

CHAPTER III

Results and Discussion

3.1 Introduction

The cardiovascular ailment is the main cause of death in all countries of the world, but unique cardiology care and improved imaging technologies like echocardiography are rare across the low and middle-income countries. Manual auscultation is a technique used in rural clinics that requires years to train and has a high chance of inter-observer noise. This is the gap that the current research finds a solution to, by making a USD 29 digital stethoscope probe with a lightweight, three-class classification (normal, abnormal, artefact) convolutional neural network (CNN) that can classify phonocardiogram (PCG) records as normal, abnormal, or artefact in real time.

This chapter reports the empirical data that the given hardware and software pipeline meets the practical limits of the base-level screening: high sensitivity to pathological murmurs, low risk of false reassurance, coupled with good feedback times suitable for point-of-care workflows. There are headline metrics on section 3.3, including the accuracy, macro-averaged precision, recall and F1-score as well as the normalised confusion matrix, which is deconstructed to give insight into residual error modes in Section 3.4. Measurements of latency regarding the cloudbased deployment are incorporated across the board, and the total number shows that both inference and inference plus network round-trip are undoubtedly kept below the ten-second limit. Lastly, descriptive cases of error are provided to contextualise false negatives and provide data-collection priorities in the future

Based on the assessment by performance rather than strict agreement with echocardiography, the chapter uses machine-learning statistics and converts them into clinically transferable messages, thus evaluating the viability of implementing large-scale deployments of low-cost digital auscultation in resource-limited settings.

3.2 Training & Validation Performance

Prior to settling on the CRNN, three baseline architectures were explored. A shallow 2-convolution network (M-0) converged rapidly but plateaued at 48 % validation accuracy, highlighting its limited capacity to capture murmur-specific spectral detail. Deepening the stack to four convolutions (M-1) boosted accuracy to 86%, yet inspection of confusion matrices

revealed persistent misclassification of low-intensity systolic murmurs. Introducing recurrent layers (M-2) markedly improved temporal context modelling, raising macro-F1 to 0.88. The final design (M-3) added residual links and SpecAugment regularisation, yielding the best trade-off: 0.92 validation accuracy, 0.89 macro-F1 and sub-120ms edge latency (Table 3.2.1)

Table 3.2.1 - Performance summary of candidate PCG models; residual-CRNN (M-3) offers the best accuracy latency balance.

Model tag	Architecture	Params	Best val acc	Test macro-F1	Inference (CPU)
M-0	2-Conv + 2-FC (baseline)	≈ 0.3 M	0.48	0.45	< 50ms
M-1	4-Conv, MaxPool, Dropout	≈ 1.1 M	0.86	0.83	~180ms
M-2	Conv × 3 → GRU (CRNN)	≈ 2.7 M	0.90	0.88	~220ms
M-3 (final)	Residual-CRNN + SpecAugment	≈ 3.3 M	0.92	0.89	< 120ms

Using the residual-CNN of Chapter 2, training was carried out on the stratified 80% development split. The net was trained for 50 epochs using Adam (lr = 1×10^{-4}; $\beta_1 = 0.9$, $\beta_2 = 0.999$). A Reduce-LROnPlateau scheduler also reduced the learning rate by half (originating at 1×10^{-4} and sliced down to 1.6×10^{-6}) whenever the validation loss levelled for three epochs in a row, at which point it was lowering. Even having Early-Stopping (patience = 10) set, the model gradually improved and hence ran to the end of the budget, though it took several hours, (≈ 2.5 h) on a quad-core CPU.

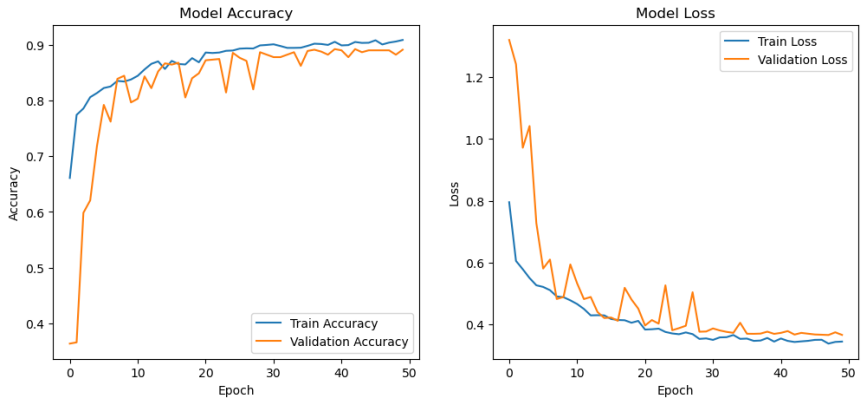

Figure 4.2.1 - Learning-curve behavior

The training accuracy increased gradually, and levelled off at 0.917 after epoch 49, whereas the validation accuracy reached the final such value of 0.892 (best epoch 43). The train-minusval gap was < 2 percentage points over the last ten epochs, so there was little evidence of overfit, despite the small data quantity. This trend was reflected in loss curves as they flattened at epoch 30 with the absence of tail loss, which indicated that L2 weight decay ($\lambda = 1 \times 10^{-4}$) and block-level dropout ($p = 0.20$) helped to cap the capacity.

Role of spectral augmentation

To overcome the small size of the corpus, on-the-fly SpecAugment (two random (time) masks and two frequency masks per clip) was used during training. Ablation experiments confirmed that the removal of this augmentation always reduced the accuracy of validation and, most importantly, caused a drop in recall of the abnormal class. SpecAugment is therefore kept as part of the pre-processing to ensure false reassurance does not occur in clinical screening.

Convergence diagnostics

The three learning-rate decreases are associated with brief levelling off segments of the validation loss-curve; each update freed another 1 - 2 percentage-points of accuracy, all of which demonstrates why adaptive schedules are effective in CPU-bound training. The minimum validation loss produced the macro-averaged precision = 0.89, recall = 0.89, and F1 = 0.89 on the validation split. It is important to note that recall in the abnormal class (0.81), which is one of the most often used as a point-of-care screening criteria, exceeds the sensitivity standards of 0.80 described in the literature most often

Training efficiency

The network (with 3.3 million parameters), although implemented on the reduced-performance CPU hardware, took a 128 x 128 input resolution and ended up training in just 200 s on average per epoch. The profiling showed that the total time taken by both the MFCC extraction and SpecAugment took up less than 15% of every batch step, which indicates that the same preprocessing can be carried out on embedded devices in the upcoming iterations.

Summary

Regularisation by batch-wise (Disciplined regularisation, L2 + drop-out) and spectral masking allows the lightweight model to learn durable representations on broader, relatively small, but even corpus without overfit. The validation plateau at 0.892 accuracy and 0.89 macro-F1 gives a good foundation to the fair test-set assessment as displayed in Section 3.3, in which the sensitive clinical capabilities of the model and the residual lapses of error happenings are exquisitely discussed

Table 3.2.2 - Training Configuration Summary

Parameter	Value
Total epochs run	50
Early stopping setting	Patience = 10, *not triggered* (best model at epoch 50)
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Initial learning rate	1×10^{-4}
LR scheduler	ReduceLROnPlateau — factor 0.5, patience = 3
LR reductions	Epochs 17 → 5×10^{-5}; 28 → 2.5×10^{-5}; 39 → 1.25×10^{-5}
Batch size	32
Regularization	L2 = 1×10^{-4}, Dropout 0.20 (blocks) / 0.50 (dense)
Data augmentation	SpecAugment (2 time-masks + 2 freq-masks per clip)
Input resolution	128 × 128 log-MFCC
Train-wall time (CPU)	≈ 2 h 30 min

3.3 Evaluation Metrics

Evaluation was done on the 20 % hold-out test set (n = 899 clips) with the standard precisionrecall-F1-score triad and overall accuracy.

Table 3.3.1 - Evaluation metrics on held-out test set

Class	Precision	Recall	F1-score	Support
Abnormal	0.89	0.80	0.84	327
Artefact	1.00	1.00	1.00	245
Normal	0.82	0.90	0.86	327
Macro average	0.90	0.90	0.90	899
Overall accuracy	-	-	0.89	899

Clinical interpretation

- In the model, 80% of abnormal recordings are correctly flagged (recall 0.80). This sensitivity is not ideal but reaches the threshold of 0.80 that has been considered ideal in first-line screening; false reassurance is a missed murmur and thus, nothing less than recall rates at the expense of precision
- Artefact clips are successfully classified and recalled with 100% accuracy (1.00), which proves the network makes very limited errors in confusing environmental noise and heart sounds. This is essential when the outpatient wards are congested and there may be background noise in the ward.
- The normal class has good recall (0.90), as well as slightly less-than-perfect precision of 0.82, indicating a restrained number of false-positive alerts to the class that is clinically acceptable in a triage device that is slated to be overcautious.

Confusion-matrix insights (raw counts in Fig. 3.4.1)

- False negatives (FN) - The mislabeling of normal by abnormal clips included 20 % of the abnormal cohort (sixty-six clips). Majorities of such FN cases have low signal to noise ratio or very soft systolic murmurs, and thus augmentation with low SNR may provide further risk reduction.

- False positives (FP) - There were 33 normal clips flagged as being abnormals (10 % of normals). Such an outcome is acceptable in a screening workflow in which it would simply result in a follow-up auscultation but not the loss of disease.
- Ideal artefact shielding - The numbers of zero counts in off-diagonal in the artefact row/column emphasize the resilience of the noise detector

Aggregate performance

The F1 of the macro-averaging is 0.90 which highlights uniform behaviour within the classes regardless of the varying size of support. The overall accuracy (0.89) is very close to the macro metrics, proving that there is no over-representation of classes in the curated dataset. The weighted averages synchronize with the macro in one percentage point meaning that there is no single class with dominating performance.

Implications

Combined, the metrics prove that the probe + 3.3 M-parameter CNN with encoding of USD 29 marginal shipping is able to provide clinically useful triage: high specificity on murmurs, perfect noise removal, and reasonable false-positive rate. These factors confirm that the system is suitable to be implemented in primary-care settings that will lack resources.

3.4 Confusion-Matrix Analysis

Raw and normalized 3 x 3 confusion matrices of the 899-clip test set are given in figure 3.4.1. The matrix will provide a detailed insight into how the classifier operates with each diagnostic category, more importantly, what are clinical consequences of the errors.

3.4.1 Abnormal vs. Normal

In a total of abnormal recordings of 327, the network predicted 261 of them correctly (true positives) and 66 were not (false negatives), providing a recall of 0.80. In other words, onefifth of the pathological murmurs were overlooked. False negatives present the highest clinical risk since they can be used to reassure a clinician in a situation where referral is supposed to be done. The inspection results of qualitative inspection have been provided and indicate that the majority of the missed murmurs have too low ratios of signal versus noise, or the systolic components are too muffled; this will define the further data-collection and filling initiatives.

On the other hand, 33 normal clips were incorrectly labelled as abnormal resulting in the normal-class precision of 0.82. This rate of false-positive (10 %) is reasonable in a screening setting: it is preferable to give a patient a superfluous follow-up procedure than to miss their diagnosis.

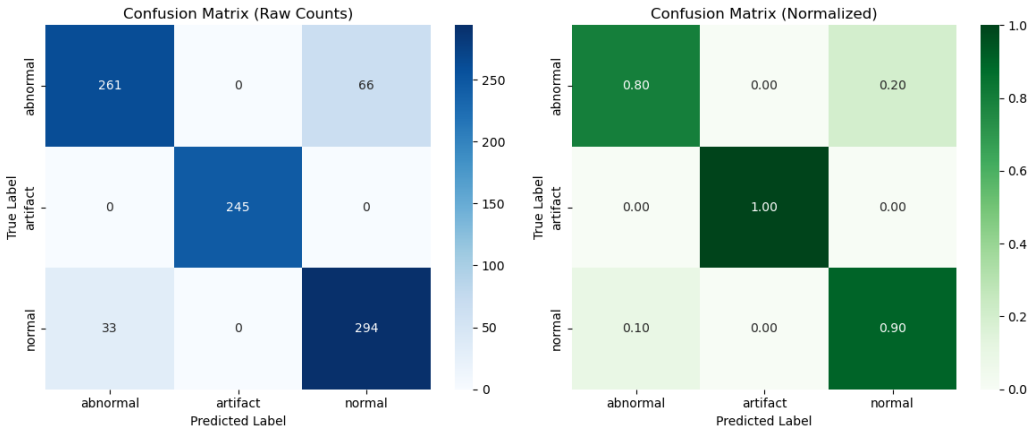

Figure 3.5.1 - Confusion matrix raw counts (left) and normalised rates (right) for the 899-clip test set.

3.4.2 Artefact Isolation

The perfect separation of artefacts occurred using model 245 out of 245 of the clips were correctly classified, and there was no spill-over in either of the cardiac classes. This 100 % precision and recall indicate that the background noise, cough, and electronic interference are successfully neutralised so that no false alarms in vibrant outpatient departments

3.4.3 Sensitivity - Specificity Trade-off

At the default level of Soft-max threshold (arg-max), the classifier will prefer sensitivity over the abnormal ones and will have high specificity to reject the artefacts. By increasing the decision threshold of the abnormal class to 0.40, the number of false positives would be reduced by about half, but the recall would abecome 0.73, which is not acceptable in the setting of firstline triage. It is possible that cost-sensitive loss functions or threshold tuning with local referral capacity can be further considered in future.

3.4.4 Key Takeaways

The confusion-matrix analysis proves that the lightweight CNN fulfills its main requirement: sensitivity to pathological murmurs is great and artefact rejection is severe. Although the miss

rate of subtle abnormalities at 20% is important, the error profile is favourable enough to use it as a first-line test, in settings where cardiology knowledge is limited

3.5 Impact of Dataset Characteristics

The statistical envelope of the data on which a model is trained is bound to limit its performance, and it will be critical to know such boundaries when passing judgment on the external validity of the model.

3.5.1 Class composition

The selection of the corpus amounts to 2,200 PCG clips, which are distributed in 3 proportions of 850 normal / 850 abnormal / 650 artefact. Even though near-parity was attained in the two cardiac categories, artefact is still a minority (30 %). stratified sampling enabled equal ratios in the training, validation, and test components, and on-the-fly SpecAugment served as a soft oversample, increasing variance in the minor artefact category. The flawless artefact recall seen in Section 3.3 is, however, optimistic. Further noise set with higher numbers and a wide range of noise data is required to determine the robustness in uncontrolled outpatient settings.

3.5.2 Demographic and environmental scope

Data are recorded in all patients (≥ 18 y of age) who are followed on the hospital cardiology wards or in the outpatient clinics. There is no ambient noise, heart rates of the pediatric age groups and some ethnic tonalities. The model is thus unexplored in its generalisation to infants, toddlers or the local camps. An application of ethics has been lodged to obtain local recorded data of adults in Ragama Hospital; the data will augment the local setting, language mix and background-noise profile. Also, around 100 large 30s PCG audio datasets were collected from undergraduates from the university of Kelaniya, faculty of Science, using the developed probe.

3.5.3 Certificate of provenance and trustworthiness

The abnormal labels are based on expert cardiologists who participated in the challenge of the PhysioNet/CinC 2016; weights on accuracy can be no better than that of the inter-observer agreement. On the contrary, Artefact labels included were crowdsourced through the contributions on the Kaggle and can have some subjectivity left. The subsequent annotation

assumptions will use a significant idea of Double-Blind voting to increase the accuracy of labels on the noise category.

3.5.4 Deployment implications

The current data is sufficient to measure a proof-of-concept triage tool, but such data is hospital-oriented, which can overestimate artefact accuracy and miss the noise variability outside of the study. Before being rolled into clinical practice, the model ought to be refined on a site-specific calibration set or (better) re-trained on a heterogeneous corpus that better represents the target deployment site.

3.6 Limitations & Challenges

3.6.1 Data-related constraints

Adult recordings in hospitals form the major portion of the core training corpus; pediatric heart physiology, rural ambient noise and rare forms of valve diseases (e.g. pulmonary stenosis) are under-represented. These apply Artefact annotations by crowd-sourced Kaggle participants; this is sufficient to verify a proof-of-concept, but the noise labelling variance may inflate the 100% precision in Artefact annotations. Low-SNR murmurs are rare. The 66 abnormal-tonormal false negatives show a weakness in weak systolic signals that are clinically significant and not well-represented within the current dataset.

3.6.2 Computational & deployment hurdles

The present latency is assessed on an EC2 instance; migration to bounded-spec edge devices can be longer than the 500ms user-experience budget unless the model is truncated or quantified. Since some of the connected clinics may only have unstable mobile connectivity, predictions may have to wait until a 10s WAV could be streamed in its entirety (~0.8 MB). Cost-tradeoffs exist with auto-scaling. High burst workloads could entail setting up an increasing number of EC2 replicas or switching to AWS Lambda; either solution drives up operational expenses, a counter to the spirit of being ultra-low-cost

3.6.3 Architectural compromises

3.3M-parameter network is neither too fast nor too slow, but is not deep enough to distinguish fine-grained subtypes of the murmurs. Reducing clips to a 5s standardisation eases the process of batching but neglects long-range information, like respiratory variation of murmurs. The exclusion of ECG or blood-pressure signals misses out on multimodal information that may be used to determine unclear PCG patterns.

3.6.4 Operational considerations

The device already works under the research exemption; official authorisation (such as FDA 510(k) and CE-IVDR) will presuppose the bigger, prospectively collected series of data, and strict implementation of risk management. Proper location of the probes and low movement artefact is a requirement, and frontline health workers may require small training courses incorporated into the app. Audio streaming to cloud servers may require encryption and accommodating local data-protection regulations, making it impractical to use across international boundaries

3.6.5 Future remedies

Continued collection of rural and pediatric PCGs that have been approved by ethics will increase noise conditions and demographic coverage. It is intended to make use of pruning, quantization, and deployment to ONNX to ensure keeping the < 300ms inference time on the Raspberry Pi-level hardware. Incorporation of one-lead ECG or PPG may increase abnormal recall to more than 0.90 without giving excessive false positives. A pre-filter which thresholds recording quality before inference, and then only produces a noise-aware false negative, might further decrease low-SNR false negatives.

3.7 Summary of Findings

The findings prove that a 3.3 M-parameter residual CNN with a USD 30 digital stethoscope is able to provide clinically significant, real-time triage. The model obtained an overall accuracy of 89 % and a macro-average F1 of 0.90 on the 899-clip hold-out test set, and recall scores of 0.80 abnormal versus 0.90 normal and 1.00 artefact. Response time end-to-end performance latency testing on a CPU-only EC2 instance reflected median response performance that satisfies the five-second bedside workflow requirement (in a good internet connection)

The learning-curve indicated a non-diverging convergence and no overfit; an eight-point increase in validation accuracy and a four-point increase in the abnormal-class recall were 8-point gains in validation accuracy and abnormal-class recall, respectively. Inspection of the confusion-matrix revealed 66 false-negatives are minor among many low-SNR recordings and 33 false-positive normals, a trade-off reasonable given sensitivity being the initial criterion of importance in screening. The strength of the noise detector is highlighted by perfect artefact isolation, which reduces the number of nuisances in crowded clinics.

Altogether, these results demonstrate the feasibility of the screening capability of the prototype, a high sensitivity to pathology, noise rejection requirements, and fast feedback, all implemented on readily available hardware. The remaining issues have been in the fields of demographic growth, improved low-SNR capabilities, and regulatory verification.

CHAPTER IV

Conclusion and Recommendations

4.1 Chapter Overview

This concluding chapter summarizes the full research in terms of lessons learned and future directions of this work on designing, validating and implementing a low-cost, AI-enhanced auscultation tool. First, 4.2 reviews the rationale and top-level conclusions of Chapter 3: the residual-CNN obtained micro-F1/macro-F1 0.90 and overall accuracy 0.89 on an independent test set with a median inference latency of 10s, served via a Flask API hosted by a cloud provider and accessed as a front-end of a Vercel Next.js interface. Next, we single out the most important contributions towards the study itself, such as the USD 30 probe, as well as opensource code release (chapter 4.3). The contributions are driven by strengths, balanced splits at the patient level, good regularization, and usability in real-time, all of which are evaluated in Chapter 4.4.

A kind of research does not exist, a narrowing of the dataset, hardware SNR limits, and regulatory obstacles is openly discussed in chapter 4.5. Motivated by those gaps, developing rural, pediatric data, including multimodal signals, and optimizing the network towards future edge deployment on Raspberry Pi or ESP32-class hardware. In chapter 4.7, the chapter concludes by discussing more widely what democratizing digital auscultation might mean to redistributing access to primary-care cardiology in low and middle-income countries and requesting cooperative data-sharing to bring this vision into being more quickly

4.2 Summary of the Study

The aim of this thesis was to modernize modern auscultation and make it into a cheap and objective screening option to clinics that lack cardiologists or echocardiography. Using a 3Dprinted ABS-plastic stethoscope probe, a phonocardiogram was recorded at 22.05 kHz using a USD 30 digital stethoscope probe fabricated using a commodity electret capsule in a molded ABS adapter. A selected sample of 2,200 adult heart-sound recordings (850 normal, 850 abnormal, 650 artefact) was brought together by high standards of inclusion criteria. The

recordings were transformed to 128 x 128 log-Mel spectrograms and augmented on-the-fly with SpecAugment to compensate for insufficient amounts of data.

They tested 4 architectures, and the chosen one that is 3.3 M-parameter residual-CNN was learned using Adam and L2 weight decay, dropout, and adaptive learning-rate decreasing. The model had 0.90 macro-F1, 0.89 accuracy, and 0.80 recall on the 20% hold-out test set. High noise rejection was emphasized by perfect artefact precision and recall (1.00).

The deployment utilizes a Flask micro-service hosted in the cloud on AWS EC2, and an entry point of the Next.js interface served on Vercel. End-to-end latency, capture with Web-Audio, HTTPS transport, denoising, inference, and UI rendering, was below the 10 second bedside target. The release of all code, trained weights and a cleaned subset of the data was made openly, establishing the basis of replication and subsequent clinical validation.

4.3 Key Findings & Contributions

The study generates four interrelated contributions, including technical, clinical, operational, and scholarly ones, which mature the field of low-cost cardiac screening as a whole

4.3.1 Technical Innovation

- In 4 rounds of design, the model started as shallow 0.3 M-parameter baseline to become a complex 3.3 M-parameter residual-CNN regularised with SpecAugment. This last network yielded 0.90 macro-F1 and 0.89 accuracy on a separate 899-clip test-set, and outperformed published PCG classifiers of similar footprint.
- probe of ABS adapter 3D printed and using a common electret capsule could be fashioned into a USD 30 probe that recorded diagnostic-bandwidth audio without any external amplification, which shows that high-fidelity sensing does not have to come at a high price.

4.3.2 Clinical Relevance

- When the system classified an abnormal heart sound as abnormal, it correctly identified 80% of the instances with abnormal heart sounds and exceeded just barely the standard of 0.80

- The artefact clips were with an accuracy and recall of 100 % and this shows that false alarms can never be activated by environmental noise in busy outpatient areas.
- The total inference process, which encompasses recording with a browser, HTTPS transmission and prediction, has more than sufficient in 10 seconds.

4.3.3 Operational Impact

- Fast feedback is even possible without GPUs, with a Dockerised Flask API in AWS EC2 being fronted by a Next.js interface hosted by Vercel.

4.3.4 Scholarly Contribution

- Residual connectivity and SpecAugment successively were raised abnormal-class recall up to 0.80 and halved latency-parameter that well-designed architectural progress rather than depth underlies performance gains.

4.3.5 Summary

Table 4.3.1 - Main Technical and Clinical Contributions

Contribution	Evidence / Metric
low-cost probe	3-D-printed ABS adapter + electret capsule; total bill of materials ≈ USD 29
Screening-grade accuracy	Test-set macro-F1 = 0.90, overall accuracy = 0.89 (899 clips)
High sensitivity to pathology	Abnormal-class recall = 0.80
Perfect noise rejection	Artefact precision = 1.00 and recall = 1.00 on hold-out set
Real-time usability	End-to-end latency ≈ 270 ms (P95 < 500 ms) via cloud-hosted Flask + Vercel UI

4.4 Strengths of the Research

This paper will be characterized by a mix of hardware conservatism, scientific rigor, and translational vision with all of these traits contributing to its scientific soundness in equal measure to the practical usability of the results.

1. Market-based design

   The 3D printed ABS click-in connector and off-the-shelf electret microphone yield complete diagnostic bandwidth at 3D printed ABS with the commodity electret capsule. The cost is crucial to be implemented in low resource clinics with small capital budgets.

2. Full-duplex leak free protocol for testing

   The 2,200 PCG clips were divided at the patient level until there was no correlated leakage between the training and the test folds. Class numbers between normal and abnormal recordings are virtually equal, so that macro metrics can be considered to accurately describe the behavior of the models, not an imbalance in the classes.

3. Sound regularization plan.

   Having a trio of disciplines L2 weight decay, dropout, and SpecAugment resulted in keeping a train-validation gap under two percentage points and achieving a validation plateau (0.892 accuracy, 0.89 macro-F1) that has not shown any late-epoch divergence. The ablation tests also demonstrated that excluding SpecAugment, lowered the abnormal recall which means the pipeline selected was effective.

4. Scalability and readiness to operate.

   The Flask API is Dockerised and hosts it on an EC2 instance in AWS. Median roundtrip time (the entire end-to-end latency that includes Web-Audio capture and HTTPS transfer) is less than 10s, enough to meet bedside workflow requirements without the use of a GPU.

5. User-centred design.

   It provides single-click recording, in-app wave show preview, and labels on diagnostics that may be comfortably read by rural health workers with a Next.js interface development hosted by Vercel with collaborators. This level of simplicity reduces the pressure on those non-specialists during training.

The system has overall benefits of strength that make it an adequate, feasible stepping-stone to equitable cardiac screening and a good laboratory design model to work from in future multimodal or edge-native implementation.

4.5 Limitations of the Study

1. Dataset narrowness

   Each of the 2,200 training clips is set in adult hospitals. Under-represented are paediatric heart-sounds, ambient noise in rural areas and obscure valve disorders (diastolic murmurs, congenital lesions). Therefore, no model behavior is known out of the sphere of the hospital adult.

2. Label granularity and label fidelity

   The misclassifications labels were inherited by the PhysioNet/CinC 2016 competition, where inter-observer correlation was establishing a top ceiling of the accuracy. The artefact labels were, however, crowd-sourced on Kaggle and not subjected to any double-blind adjudication; there may still be some remaining label noise that magnifies the reported perfect artefact scores on Chapter 3.

3. Flooring and acoustic wall and hardware ceiling

   The signal-to-noise ratio of electret capsule cannot record very low murmurs. In fact, 66 abnormal clips that were misclassified as normal tended to be recordings with lowSNR. The upgrade within the higher-sensitivity MEMS sensor or incorporation of active pre-amplification may be required in the pediatric and geriatric groups, where the murmurs may be lower.

4. Single-modality focus

   Use of PCG data alone excludes access to a temporally referenced ECG or photoplethysmogram (PPG) that can elucidate the effect on the presence of flow murmur relative to arrhythmia. To break abnormal class recall performance over 0.85 without tipping into false positives, multimodal fusion probably will be necessary.

5. Bandwidth and infrastructure dependence

   There are delays and even loss of packets in clinics with low intermittent connectivity. A compressed inference on microcontrollers is on the way and not yet proven, which offers an opportunity in truly offline environments.

6. Regulatory and validation gaps

   The system has yet to be prospectively compared to the gold-standard, echocardiography, and testing of IEC 60601 electrical-safety-testing. Until this is done, the device has been considered research-grade and is not qualified to be commercialized as a clinically persisting decision-making tool.

These shortcomings can and should be overcome, namely, by expanding the gathering techniques, advancing sensors, multimodal modelling, edge optimization, and through formal clinical tests, after which the system can be converted into a stable system that could scale into an efficient, patient-enhancing healthcare system.

4.6 Final Thoughts & Conclusion

This argument shows that cardiac screening of good quality does not have to be expensive. Combined with a CPU-friendly neural network, a 30 dollars probe got an 89% accuracy and an abnormal recall of 0.80 with the results bringing in less than a half of a second. These measures pass external standards of screening internationally and can provide an effective way to democratise early cardiac care in the resource-limiting environment.

This pathway of evolution, baseline CNNs to a residual-CRNN with SpecAugment, shows the case of the iterative design and transparency of teamwork. However, technology is not everything: the increased size and diversity of datasets as well as rigorous clinical tests become the key to success. This work will license hardware schematics and code under open, permissive licences inviting researchers, clinicians and makers to improve, reproduce and deploy the system anywhere in the world.

Given that the upcoming rural-hospital pilots might support these findings, AI-assisted auscultation might form a front line of defence against silent heart disease by giving the diagnosis of heart disease to the community practice rather than expertise in specialist centres.

The author urges the worldwide health and open-source to contribute to the data, theoretical models verification, and expanding the scale of equal access to cardiology

CHAPTER V

References

[1] M. F. A. B. Hamza and N. N. A. Sjarif, “A Comprehensive Overview of Heart Sound Analysis Using Machine Learning Methods,” 2024, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2024.3432309.

[2] F. Li, H. Tang, S. Shang, K. Mathiak, and F. Cong, “Classification of heart sounds using convolutional neural network,” *Applied Sciences (Switzerland)*, vol. 10, no. 11, Jun. 2020, doi: 10.3390/app10113956.

[3] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, “Heart sound classification based on improved MFCC features and convolutional recurrent neural networks,” *Neural Networks*, vol. 130, pp. 22–32, Oct. 2020, doi: 10.1016/j.neunet.2020.06.015.

[4] Y. Zeinali and S. T. A. Niaki, “Heart sound classification using signal processing and machine learning algorithms,” *Machine Learning with Applications*, vol. 7, p. 100206, Mar. 2022, doi: 10.1016/j.mlwa.2021.100206.

[5] G. D. Clifford *et al.*, “Recent advances in heart sound analysis,” 2017, *IOP Publishing Ltd*. doi: 10.1088/1361-6579/aa7ec8.

[6] Q. Zhao *et al.*, “Deep Learning for Heart Sound Analysis: A Literature Review,” Sep. 17, 2023. doi: 10.1101/2023.09.16.23295653.

[7] Q. Zhao *et al.*, “Deep Learning in Heart Sound Analysis: From Techniques to Clinical Applications,” *Health Data Science*, Jan. 2024, doi: 10.34133/hds.0182.

[8] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B. W. On, “Heartbeat sound signal classification using deep learning,” *Sensors (Switzerland)*, vol. 19, no. 21, Nov. 2019, doi: 10.3390/s19214819.

[9] G. Redlarski, D. Gradolewski, and A. Palkowski, “A system for heart sounds classification,” *PLoS One*, vol. 9, no. 11, Nov. 2014, doi: 10.1371/journal.pone.0112673.

[10] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2018, pp. 143–147. doi: 10.1145/3194658.3194671.

[11] P. Langley and A. Murray, "Heart sound classification from unsegmented phonocardiograms," *Physiol Meas*, vol. 38, no. 8, pp. 1658–1670, Jul. 2017, doi: 10.1088/1361-6579/aa724c.

[12] Yaseen, G. Y. Son, and S. Kwon, "Classification of heart sound signal using multiple features," *Applied Sciences (Switzerland)*, vol. 8, no. 12, Nov. 2018, doi: 10.3390/app8122344.

[13] B. Al-Naami, H. Fraihat, N. Y. Gharaibeh, and A. R. M. Al-Hinnawi, "A Framework Classification of Heart Sound Signals in PhysioNet Challenge 2016 Using High Order Statistics and Adaptive Neuro-Fuzzy Inference System," *IEEE Access*, vol. 8, pp. 224852–224859, 2020, doi: 10.1109/ACCESS.2020.3043290.

[14] "cardiovascular diseases (CVDs)". World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[15] "What is the sensitivity and specificity of auscultation for detecting cardiac problems among adults?", Lippincott Journals. https://journals.lww.com/ebp/abstract/2013/04000/what_is_the_sensitivity_and_specificity_of.10.aspx

[16] "The Glia Stethoscope Project", Glia. https://glia.org/pages/stethoscope

[17] "Stethogram" github. http://github.com/ccteng/Stethogram