PULSECHECK – LIGHTWEIGHT ML BASED CLASSIFICATION OF HEART SOUNDS USING PCG SIGNALS

A SOCIALLY REVELANT MINI PROJECT REPORT

Submitted by

VEDHA SHREE R [211423104722]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(AnAutonomous Institution, Affiliated toAnna University, Chennai)

OCTOBER 2025

BONAFIDE CERTIFICATE

Certified that this project report “PULSECHECK – LIGHTWEIGHT ML BASED CLASSIFICATION OF HEART SOUNDS USING PCG SIGNALS” is the bonafide work of VEDHA SHREE R(211423104722) who carried out the project work under my supervision.

Signature of the HOD

Dr L. JABASHEELA, M.E., Ph.D., PROFESSOR AND HEAD, Department of CSE

Panimalar Engineering College,

Chennai - 123

Signature of the Supervisor Dr.V.SUBEDHA,M.Tech., Ph.D., PROFESSOR,

Department of CSE

Panimalar Engineering College,

Chennai- 123

Submitted for 23CS1512 - Socially Relevant Mini Project Viva-Voce Examination held

on...........................

INTERNAL EXAMINER EXTERNAL EXAMINER

ii

DECLARATION BYTHE STUDENT

I VEDHASHREE R(211423104722) hereby declare that this project report

titled “PULSECHECK – LIGHTWEIGHT ML BASED CLASSIFICATION OF

HEART SOUNDS USING PCG SIGNALS” under the guidance of

Dr.V.SUBEDHA M.Tech., PH.D., is the original work done by me and I

have not plagiarized or submitted to any other degree in any university

by me.

SIGNATURE OFTHE STUDENT

VEDHASHREE R(211423104722)

iii

ACKNOWLEDGEMENT

I express my deep gratitude to our respected Secretary and Correspondent Dr.P.CHINNADURAI, M.A., Ph.D., for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

I would like to express my sincere thanks to our Directors

Tmt.C.VIJAYARAJESWARI, Dr.C.SAKTHIKUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHIKUMAR B.E.,M.B.A.,Ph.D., for providing me

with the necessary facilities for completion of this project.

I also express my gratitude to our Principal Dr.K.MANI, M.E., Ph.D., for his timelyconcern and encouragement provided to us throughout the course.

I thank the HOD of CSE Department, Dr.L.JABASHEELA, M.E., Ph.D., for the support

extended throughout the project.

I would like to thank our Project Coordinator and Project Guide Dr.V.SUBEDHA, M.Tech., Ph.D., and all the faculty members of the Department of CSE for their advice and suggestions for the successful completion of the project.

I also extend my heartfelt thanks to all the faculty members of the Department of Computer

Science and Engineering for their encouragement and advice, which greatly contributed to

the successful completion of my project.

VEDHASHREE R (211423104722)

iv

ABSTRACT

Cardiovascular diseases are one of the major causes of deaths worldwide,

which highlights the need for accessible and affordable screening

techniques. Traditional methods using a stethoscope is subjective and it also

requires clinical expertise, which makes early detection a challenge in rural

areas.

This project, Pulsecheck - Lightweight Ml Based Classification Of Heart

Sounds Using Pcg Signals, aims to provide machine learning- based system

for classification of heart sounds into categories like Normal and Abnormal.

The system uses Phonocardiogram (PCG) signals available publicly as

datasets. These signals undergoes preprocessing steps which include

resampling, noise reduction, and normalization to get consistent high quality

data. The features are extracted from the signals and are sent to a lightweight

Machine Learning pipeline based on SMOTE for balancing class and

classification using XGBoost. The experimental evaluation shows that

PulseCheck provides high accuracy in detecting categories like abnormal

and normal heart sounds making it efficient and suitable for integration into

mobile applications, telemedicine platforms and digital stethoscopes. The

system can help as a screening tool for healthcare providers thereby

contributing to early diagnosis and prevention of cardiac diseases.

v

TABLE OF CONTENTS

CHAPTER NO. TITLE PAGE NO.

ABSTRACT v

LIST OF FIGURES ix

LIST OFABBREVIATIONS x

LIST OF TABLES xi

1.

1.1

1.2

1.3

1.4

2.

2.1

2.2

2.3

2.4

2.5

2.6

3.

3.1

INTRODUCTION

Problem Definition

Objectives

Significance of the Study

Project Scope

LITERATURE SURVEY

Introduction

Conventional Machine LearningApproaches

Signal Processing and Transform-Based Methods

Ensemble andAdvanced Machine Learning Models

Deep Learning Approaches

Research Gaps

SYSTEMANALYSIS

Existing System

1

2

2

3

3

5

5

5

7

9

10

11

13

13

vi

3.2 Proposed System 14

3.3 Feasibility Study 17

3.4 ComparativeAnalysis of PulseCheck 18

4. SYSTEM DESIGN 20

4.1 SystemArchitecture Diagram 20

4.2 Data Flow Diagram 20

4.3 Use Case Diagram 21

4.4 Class Diagram 22

4.5 Sequence Diagram 23

4.6 Activity Diagram 24

5. SYSTEMARCHITECTURE 26

5.1 Introduction 26

5.2 System Overview 26

5.3 Module Design Specification 27

5.4 Input Design 33

5.5 Output Design 33

6. SYSTEM IMPLEMENTATION 35

6.1 Introduction 35

6.2 Implementation Workflow 35

6.3 Sample Coding 36

6.4 Integration of Modules 41

6.5 Real-Time Deployment Considerations 41

vii

7. SYSTEM TESTING 42

7.1 Introduction 42

7.2 Testing Strategy 42

7.3 Test Cases 44

7.4 Observation and Discussion 45

8 CONCLUSION AND FUTURE WORK 46

8.1 Conclusion 46

8.2 Limitations 47

8.3 Future Work 47

9 APPENDICES 48 A1 - SDG GoalAlignment 48 A2 - Sample Screenshots 49 A3 - Paper Publication 50 A4 - Plagiarism report 58

10 REFERENCES 71

viii

LIST OF FIGURES

FIGURE NO.

1.1

2.1

3.2

4.1

4.2

4.3

4.4

4.5

4.6

5.1

A2.1

A2.2

A2.3

A2.4

A2.5

FIGURE DESCRIPTION

Block Diagram of PulseCheck System

MLAlgorithm accuracy of Heart Sound class

Proposed SystemArchitecture

System Architecture diagram

Data Flow Diagram

Use Case Diagram

Class Diagram

Sequence Diagram

Activity Diagram

Conceptual SystemArchitecture Diagram

Launch Page

File Upload Interface

Classification Ouput

Prediction Probabilities

MFCC Visualization

PAGE NO.

5

13

19

31

33

35

36

37

38

40

46

47

48

49

50

ix

LIST OFABBREVIATIONS

S. NO ABBREVIATIONS

1 PCG - Phonocardiogram

2 ML – Machine Learning

3 SMOTE – Synthetic Minority Oversampling Technique

4 MFCC – Mel-Frequency Cepstral Coefficients

5 XGBoost – Extreme Gradient Boosting

6 CNN – Convolutional Neural Network

7 SVM – Support Vector Machine

8 DWT – Discrete Wavelet Transform

9 STFT – Short-Time Fourier Transform

10 RNN – Recurrent Neural Network

11 DNN – Deep Neural Network

12 AI – Artficial Intelligence

13 WHO – World Health Organization

14 SDG – Sustainable Development Goals

15 GUI – Graphical User Interface

16 ROC – Receiver Operating Characteristic

17 API –Application Programming Interface

18 CSV – Comma-Separated Values

19 PDF – Portable Document Format

x

LIST OF TABLES

TABLE NO.

5.1

5.2

5.3

5.4

5.5

6.1

7.1

7.2

7.3

8.1

TABLE DESCRIPTION

Preprocessing Steps

Class Distribution Before and After SMOTE

Sample Output Table

Input Design Specifications

Output Design Specifications

Sample Output

Unit Testing Results

Integration Testing Scenarios

Performance Testing Metrics

Achievements vs Objectives

PAGE NO.

28

30

32

33

34

39

42

43

44

76

xi

CHAPTER 1

INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide,

accounting for approximately 17.9 million deaths annually, which is nearly 32% of all global deaths (World Health Organization, 2025). Early detection of CVDs is crucial for effective management and treatment.

Traditional diagnostic methods, primarily using a stethoscope, rely heavily on the

clinician’s skill and experience. Manual auscultation is prone to human error, particularly in noisy environments, or when subtle heart abnormalities, such as systolic and diastolic murmurs, are present. This makes early detection difficult, especially in rural or resource-constrained areas where access to cardiologists is limited.

Phonocardiogram (PCG) signals provide a digital recording of heart sounds, which

can be analyzed computationally. With advances in machine learning (ML), these signals can now be classified automatically,= distinguishing between normal and abnormal heart sounds.

PulseCheck is designed as a lightweight ML-based system to classify PCG signals

efficiently, enabling real-time, portable, and accessible cardiac sc reening. The system integrates preprocessing, feature extraction, class balancing, and classification, offering a deployable solution for telemedicine, mobile health applications, and smart stethoscopes.

1

1.1 PROBLEM DEFINITION

The traditional method of heart sound analysis suffers from several key limitations:

1. Subjectivity – Manual auscultation depends on clinician expertise and auditory accuracy.

2.Prone to Errors – Environmental noise and subtle murmurs can lead to misdiagnosis.

3.Resource Dependence – Deep learning-based computerized systems often require large datasets and high computational power, limiting deployment in mobile or low-resource settings.

Problem Statement:

Design a system that is accurate, lightweight, and deployable, capable of automatically classifying heart sounds from PCG signals as Normal orAbnormal, while being suitable for real-time use on portable devices.

1.2 Objectives

The key objectives of the PulseCheck project are:

 Preprocess PCG signals through resampling, noise reduction, and normalization to obtain clean, consistent inputs.

 Extract effective features from the signals, including time-domain, frequency-domain, and MFCC-based features.

 Address class imbalance using SMOTE (Synthetic Minority Oversampling Technique) to improve sensitivity to abnormal heart sounds.

 Train and evaluate an XGBoost classifier for accurate classification of Normal andAbnormal heart sounds.

 Provide a fast, portable, and clinically applicable solution for early cardiac screening.

1.3 Significance of the Study

The PulseCheck system addresses several real-world challenges:

 Accessibility – Provides a low-cost, portable tool for early cardiac screening in rural and low-resource areas.

 Efficiency – Reduces dependency on highly skilled clinicians for initial diagnosis.

 Integration – Can be integrated into mobile health applications, telemedicine platforms, or smart stethoscopes.

 Early Detection – Helps identify abnormal heart sounds before severe cardiovascular events, enabling timely intervention.

1.4 Project Scope

The scope of PulseCheck includes:

ApplicationArea

Healthcare Facilities

Telemedicine

Smart Devices

Research

Description

Hospitals and clinics can use the system for early screening of heart abnormalities.

Remote monitoring of patients with real-time heart sound classification.

Integration into AI-enabled stethoscopes for portable, low-resource diagnostics.

Benchmarking on public datasets and exploration of lightweight ML methods in biomedical signal processing.

3

Boundaries of the Project:

 The system is limited to binary classification (Normal/Abnormal).

 It is not a replacement for clinical diagnosis but serves as a screening tool.

 heart cond

Fig.1.1 Block Diagram of PulseCheck System

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

The Cardiovascular disease (CVD) is the largest global cause of death, estimated at 17.9 million annually, based on the World Health Organization (WHO). That is close to 32% of all global mortality, and a large part of it is due to underdiagnosed or un treated disease. The earlier the CVD is discovered and treated, the larger the decrease in mortality, morbidity, and long-term healthcare expenditures. But in most low-resource and rural environments, cardiologists, echocardiography, and sophisticated diagnostic units are not available. Low-cost, transportable, mechanized diagnostic equipment for screening the people at the grassroot level must be innovated. PCG—the stethoscope or digital sensor recording of cardiac sounds—is one among the numerous clinical markers present for the diagnosis of CVD.

Sounds of S1 (first heart sound), S2 (second heart sound), and murmurs give useful information about cardiac function, valve disease, and blood flow. Clinicians have conventionally performed manual auscultation, whereby the clinician depends on hearing acuity to detect abnormalities. The process is subjective, dependent only on clinician expertise, and usually of no assistance in noisy situations. In addition, underlying abnormalities like systolic and diastolic murmurs may escape detection with traditional inspection. Computerized PCG signal analysis is a preferable alternative, though.

2.2 Conventional Machine LearningApproaches

Early studies were centered around feature-based machine learning models, in which domain specialists extracted manually time-domain and frequency-domain features from PCG signals. Popular features were Mel-Frequency Cepstral Coefficients (MFCCs), spectral energy, zero-crossing rate, and statistical descriptors like mean and variance.

For instance, Maknickas and Maknickas (2017) [3] showed that the integration of

MFCC features with classifiers such as Support Vector Machines (SVM) was capable 5

of distinguishing between normal and abnormal heart sounds. Likewise, Potes et al. (2017) [2] investigated an ensemble system consisting of feature-based and machine learning classifiers, which presented better robustness compared to single-model systems.

These works emphasized the role of precise feature engineering in delivering consistent classification performance.

2.2.1 Studies and Findings

Maknickas & Maknickas (2017) [3]

o Dataset: PhysioNet/CinC Challenge; 1,000 PCG recordings

o Features: MFCCs

o Algorithm: SVM classifier

o Results: 88% accuracy

o Insights: MFCCs effectively capture spectral characteristics of heart sounds.

o Limitations: Sensitive to noise; requires careful feature engineering.

Potes et al. (2017) [2]

o Dataset: CinC Challenge

o Features: MFCCs + time-domain statistical features

o Algorithm: Ensemble of SVM and Random Forest

o Results: 90% accuracy

o Insights: Ensemble complementary

approach improves robustness; combines classifiers.

o Limitations: Computational complexity increases with ensemble.

Kumar & Singh (2022) [5]

o Dataset: Local PCG recordings (500 normal, 200 abnormal)

o Features: Spectral energy, entropy, statistical measures

o Algorithm: Random Forest

o Results: 87% accuracy

o Insights: Simpler ML models can achieve reasonable accuracy on small datasets.

o Limitations: Poor performance on noisy signals and imbalanced datasets.

Observations:

Conventional ML methods are computationally light, interpretable, and suitable for small datasets. However, they rely heavily on expert-crafted features and are sensitive to noise and imbalance.

2.3 Signal Processing and Transform-Based

Methods A number of works have utilized time-frequency analysis methods to acquire more elaborate representations of heart sounds. Wavelet Transform (WT), Short-Time Fourier Transform (STFT), and Discrete Wavelet Transform (DWT) have been frequently used to record both spectral and temporal changes in PCG signals. For example, Kay and Mathews (2020) [4] employed wavelet-based feature extraction with machine learning algorithms to process PCG recordings. They were able to demonstrate that the use of multi-resolution analysis led to increased sensitivity toward subtle abnormalities like heart murmurs, which are commonly overlooked in strictly time domain processing. Likewise, Roy and Das (2022) [6] used a hybrid set of MFCC and DWT features and illustrated that hybrid sets offer greater discriminability compared to single feature solutions.

7

2.3.1 Studies

Kay & Mathews (2020) [4]

o Technique: Wavelet-based feature extraction

o Dataset: PhysioNet recordings

o Results: Enhanced detection of subtle murmurs

Insight: Multi-resolution analysis identifies high-frequency oscillations indicative of abnormal heart sounds

Roy & Das (2022) [6]

o Technique: MFCC + DWT hybrid features

o Dataset: Local PCG recordings

o Results: Hybrid features outperform single-feature models

o Insight: Combining spectral and temporal information increases classifier discriminability

Verma & Sharma (2022) [9]

o Technique: STFT-Mel-WSST based CNN

o Results: Improved classification using spectral feature learning

o Insight: Transform-based input allows deep models to capture high-resolution patterns

Advantages:

o Detects subtle abnormalities like systolic/diastolic murmurs.

o Improves robustness to variations in heart sound morphology.

Limitations:

o Feature extraction is computationally expensive.

o Complexity may hinder real-time deployment.

2.4 Ensemble andAdvanced Machine Learning

Models As PCG datasets have become more complex, ensemble models and more sophisticated classifiers have become the central focus. XGBoost, Random Forest, and ensemble feature fusion methods have been found to efficiently deal with noisy and imbalanced datasets.

Ahmed and Sinha (2023) [8] presented a mutual information-based hybrid feature fusion technique that enhanced classification accuracy by synergistically combining complementary features. Likewise, Reddy and Das (2022) [7] utilized wavelet-based ensemble models for effective detection of abnormal heart sounds with high accuracy over various datasets. These findings and results emphasize the advantage of combining multiple classifiers and features to improve performance under real-world scenarios.

2.4.1 Studies

Ahmed & Sinha (2023) [8]

o Method: Mutual information-based hybrid feature fusion o Features: MFCC + spectral energy

o Results: 92% accuracy

o Insight: Fusion of complementary features improves model generalization

Reddy & Das (2022) [7]

o Method: Wavelet-based ensemble deep learning o Results: 91% accuracy

o Insight: Ensemble models handle noise and imbalance better than single Green & White (2021) [11]

o Method: MFCC ensemble o Results: 90% accuracy

o Insight: Aggregating multiple models improves prediction stability

Strengths:

 Handles noisy, imbalanced datasets 9

 Improves generalization and robustness

Weaknesses:

 Computationally intensive.

 Dependent on quality of feature extraction.

2.5 Deep Learning Approaches

Recent research has focused on deep learning models for heart sound classification. CNNs and RNNs are specifically efficient at automatically learning hierarchical features from raw or spectrogram-transformed signals. For instance, Takahashi and Sun (2022) [10] introduced a ConvTransformer encoder for detecting local and global features in heart sound data. Deep learning techniques, though tend to have state-of-the-art accuracy, are computationally demanding and need large datasets, which can constrain their application in low-resource settings. This has spurred efforts towards hybrid methods that blend deep learning feature extraction with lightweight classifiers.

2.5.1 Studies

Takahashi & Sun (2022) [10]

o Model: Conv-Transformer o Dataset: PhysioNet

o Results: 94% accuracy

o Insight: Captures both local and global patterns in heart sounds

Maknickas & Maknickas (2017) [3]

o Model: CNN on MFCC input o Results: 91% accuracy

o Insight: Convolutional layers capture hierarchical spectral features

Kaur & Gill (2021) [12]

o Model: Deep ANN + HPSS o Results: 92% accuracy

o Insight: Harmonic-percussive separation improves feature representation

Xu & Zhao (2023) [14]

o Model: Explainable CNN o Results: 93% accuracy

o Insight: Provides interpretability using attention maps

Strengths:

o High accuracy and automated feature extraction. o Effective for complex or subtle patterns.

Limitations:

o Requires large datasets and high computational power.

o Limited interpretability without explainableAI techniques. o Hard to deploy on low-resource devices

2.6 RESEARCH GAPS

Based on the literature review, the following gaps have been reported:

1. High Computational Cost – Deep models, though being accurate, are computationally costly and hard to implement on smartphones.

2. Data Imbalance – Most PCG datasets are overwhelmed by normal heart sounds, and therefore, there is class bias in classifiers.

3. Sensitivity to noise – Classifiers built with clean signals perform badly in noisy environments or real-life situations.

4. Limited Accessibility implies that existing systems are not designed to be interfaced with low-power devices or telemedicine platforms.

In spite of such advancements, a number of challenges remain in automatic PCG classification:

1. Noise and Artifacts: Atmospheric noise, respiration, and patient motion can contaminate PCG signals and decrease classification accuracy.

2. Class Imbalance: Abnormal heart sounds are usually underrepresented in datasets, thus leaning models toward the normal class.

3. Model Interpretability: Deep learning models, although precise, tend to be "black boxes," and it is hard to justify predictions to clinicians.

11

4. Real-Time Deployment: Most state-of-the-art models consume high ostic

Fig.2.1.MLAlgorithm accuracy for heart sound class

CHAPTER 3

SYSTEMANALYSIS

3.1. Existing System

In the traditional healthcare environment, diagnosis of cardiovascular disease using heart sounds is carried out by auscultation manually with the aid of a stethoscope. Phonocardiogram (PCG) signal analysis largely relies on the physician's experience and hence is subjective and prone to human error. There have been some efforts to classify using traditional ML models (Random Forest, k-NN, SVM) and deep learning models (RNNs, CNNs). While CNNs and RNNs improve accuracy by learning on raw or spectrogram-extracted PCG signals, they:

o Require large data for training.

o Require GPU/TPU for real-time inference.

o Are not suitable for low-resource or mobile environments.

o Current methods lack either accuracy and hardness (traditional ML) or performance and deployability (deep learning).

3.1.1 Existing ComputationalApproaches

1. Conventional Machine Learning

Classifiers like SVM, k-NN, and Random Forest rely on manually extracted features (MFCC, spectral energy, zero-crossing rate).

Pros: Lightweight and interpretable.

Cons: Sensitive to noise, feature-dependent, lower accuracy in complex cases.

2. Deep Learning Models

13

CNNs and RNNs automatically extract features from raw or spectrogramtransformed PCG signals.

Pros: High accuracy, capable of detecting subtle abnormalities.

Cons: Require large datasets, computational resources (GPU/TPU), and often act as black boxes.

Observation: Existing methods either lack accuracy in real-world noisy data (traditional ML) or are resource-intensive (deep learning), limiting deployability on mobile or handheld devices.

3.2 Proposed System

The PulseCheck system is a deployable and lightweight machine learning architecture aimed for automatically classifying phonocardiogram (PCG) signals into either Normal or Abnormal. The system is constructed to optimize accuracy, performance, and resource utilization, rendering it viable for usage in telemedicine, handheld diagnostic, and smart stethoscope applications. PulseCheck does this through a well-crafted pipeline of preprocessing, feature extraction, class balancing, classification, and output generation that supports solid performance against noisy or class-imbalanced datasets.

3.2.1 Preprocessing

Heart sound recordings might be quite different due to environmental noise, patient movements, and variability between recording devices. For the treatment of these challenges, the initial process of PulseCheck concentrates on signal cleaning and standardization.

All PCG recordings are initially resampled into a unified frequency, so that the resolution in time remains uniform across recordings. This is important since varying sampling rates may warp both time-domain as well as frequency domain characteristics, resulting in incorrect classification.

Then, a Butterworth band-pass filter is used to eliminate low frequency noise from motion artifact and high-frequency interference that can be generated by the recording environment. The filtering keeps the vital contents of the heart sounds, including the first and second heart sounds (S1 and S2) and any possible murmurs, while rejecting noncontributory signal content.

Lastly, the signals are normalized to normalize the amplitude between recordings. This is to ensure that differences in recording volume or sensitivity of the stethoscope do not affect the classifier's output. At the completion of preprocessing, the PCG signal in each case is clean, standardized, and ready for feature extraction.

3.2.2. Feature Extraction

Feature extraction is a key process since it converts unprocessed PCG signals into measurable representations that can be interpreted by a machine learning model. PulseCheck uses both time-domain and frequency-domain features in tandem to compile the key properties of heart sounds.

Among the used features, Mel-Frequency Cepstral Coefficients (MFCCs) are particularly responsible for capturing the spectral characteristics of the signal, highlighting the fine differences between the normal and abnormal heart sounds. The distribution of signal power over frequency bands is quantified by spectral energy, which is especially effective in detecting murmurs or irregular heartbeats with increased energy in certain ranges.

Other characteristics, including entropy, measure signal complexity, assisting in segregating regular and irregular cardiac activity. The rate of zerocrossing counts the number of times the signal passes the zero amplitude line and serves to provide further detail on fast oscillations or high frequency elements that are found within anomalous signals.

15

Lastly, statistical attributes like mean, variance, skewness, and kurtosis are derived from both the energy and amplitude of the waveform. The attributes complement their spectral counterparts, describing general patterns in the morphology of the heart sound. Combined, the two ensure that PulseCheck is able to identify both high-resolution spectral changes and general statistical patterns in the data.

Fig.3.2.Proposed SystemArchitecture

3.2.3. Class Balancing

Medical data frequently exhibits class imbalance, as healthy heart sounds are far more prevalent than abnormal ones. This class imbalance has the effect of skewing a model such that it is less responsive to rare but clinically significant abnormal instances. To combat this, PulseCheck uses SMOTE (Synthetic Minority Oversampling Technique), which creates artificial representatives of the minority class by interpolating between current abnormal instances.

By balancing the classes, the system enhances generalization and minimizes false negatives, something that is very important in the case of a screening tool that focuses on early detection of cardiac problems.

3.2.4. Lightweight Classification

The core of PulseCheck's classification potential is contained within XGBoost, a gradient-boosted decision tree classifier. XGBoost is selected due to its high accuracy, efficiency, and interpretability. It has the ability to deal with large sets of features and can detect complicated nonlinear relationships among the features. Moreover, the model enables feature importance extraction, where clinicians are able to determine which features of the PCG signal played the most significant role in a classification.

In order to make the model run as efficiently as possible, GridSearchCV is used to tune the hyperparameters. Hyperparameters like the number of estimators, the depth of the trees, learning rate, and subsampling ratios are systematically evaluated using stratified 5-fold cross validation. This precise tuning ensures that the resulting model is both precise and stable on various subsets of the data

3.2.5. Output

Following classification, PulseCheck generates a binary output, i.e., Normal or Abnormal, for the input PCG signal. The low-latency prediction capability of the system makes it adequate for deployment in real-time clinical or remote monitoring applications. Its low-weight nature permits deployment on telemedicine platforms, handheld stethoscopes, and other intelligent diagnostic devices for aiding early screening and timely intervention of cardiovascular disease.

By combining judicious preprocessing, thorough feature extraction, class balancing, and optimized XGBoost classification, PulseCheck offers a feasible, efficient, and clinically pertinent solution for automated heart sound analysis.

3.3 FEASIBILITY STUDY

The feasibility study was carried out to determine whether the system under consideration is feasible and practicable.

3.3.1.TECHNICALFEASIBILITY

 Tools: Python, Librosa, Scikit-learn, XGBoost.

 Dataset: PhysioNet/CinC Challenge (3,240 PCG signals).

17

 Hardware: Could run on standard laptops without GPU, ideal for deployment in low-power environments.

3.3.2.ECONOMIC FEASIBILITY

 Dataset is free and available openly.

 Takes only modest computational power.

 Affordable in comparison to current medical equipment.

3.3.3.OPERATIONALFEASIBILITY

 Easy to use and needs minimal technical expertise to deploy.

 Can be integrated into healthcare processes, mobile health, and electronic stethoscopes.

 Helps in workload reduction of physicians by supporting decision-making.

3.3.4.LEGALFEASIBILITY

 Make use of publicly available dataset with proper licensing.

 Complies with ethical requirements of medical research and deployment.

3.3.5 SCHEDULE FEASIBILITY

 Project can be done in a 3–4 month timeframe, with data preparation, training, evaluation, and documentation.

3.4 COMPARATIVEANALYSIS OF PULSECHECK

Feature

Accuracy Computation Data Requirement Deployability

Interpretability

Traditional ML

Moderate Low Low Moderate

High

Deep Learning

High High High Low

Low

PulseCheck

High Low Moderate High

High

Class Imbalance Handling Limited Moderate Yes (SMOTE)

Observation:

PulseCheck balances accuracy, efficiency, interpretability, and deployability, addressing gaps in previous approaches.

19

CHAPTER 4

SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE DIAGRAM

Definition:

Asystem architecture diagram depicts the high-level structure of the system. It illustrates how the major components (data acquisition, preprocessing, feature extraction, classification, and output generation) interact with each other.

Purpose:

To show the modular design of PulseCheck and the flow of data between layers such as input, processing, and output.

Fig.4.1. SystemArchitecture Diagram 4.2 Data Flow Diagram Definition:

AData Flow Diagram (DFD) represents how information moves through the system. It highlights the flow of data between external entities, processes, and data stores.

Purpose:

To visualize the logical movement of PCG data—from acquisition to classification and result generation.

4.3 Use Case Diagram

Definition:

AUse Case Diagram defines the interaction between users (actors) and the system’s functional modules.

Purpose:

To describe how end-users such as patients and doctors interact with PulseCheck— uploading heart sounds, viewing results, and downloading diagnostic reports.

21

Fig.4.3.Use Case Diagram

4.4 Class Diagram

Definition:

AClass Diagram describes the object-oriented structure of the software. It shows the classes, attributes, methods, and relationships between them.

Purpose:

To outline the main SignalProcessor, FeatureExtractor, C

4.5 Sequence Diagram

Definition:

ASequence Diagram displays the chronological sequence of interactions between system components during operation.

Purpose:

To illustrate the o —showing how modules commun

Fig.4.5. Sequence Diagram

4.6 Activity Diagram

Definition:

AnActivity Diagram visualizes the dynamic flow of actions and decisions within the system. It resembles a flowchart showing conditional branches and parallel flows.

Purpose:

To depict the logical flow of operations in PulseCheck, starting from data upload to classification and display of results.

23

Fig.4.6.Activity Diagram

CHAPTER 5

SYSTEM ARCHITECTURE

5.1 Introduction

The system architecture represents the structural design of PulseCheck, a machine

learning-based heart sound classification system. It ensures that the system processes

phonocardiogram (PCG) signals efficiently and produces accurate predictions in

realtime, low-resource environments.

PulseCheck is designed as a modular, scalable, and maintainable system to allow easy

integration with mobile devices, telemedicine platforms, andAI-enabled stethoscopes.

The architecture emphasizes:

 Accuracy and reliability of heart sound classification.

 Lightweight computation suitable for handheld or mobile deployment.

 Clear data flow for maintainability and system understanding.

 Flexibility for future enhancements, such as multi-class classification

or integration with other biomedical signals.

5.2 System Overview

PulseCheck comprises five major modules, each responsible for a specific part of the

system pipeline:

 Data Acquisition Module – Collects PCG signals from datasets or real-time

recording devices.

 Preprocessing Module – Cleans and standardizes signals for analysis.

 Feature Extraction Module – Converts PCG signals into a structured feature set.

25

 Classification Module – Uses XGBoost to classify signals as Normal orAbnormal.

 Output Module – Generates results in a readable, actionable format.

Figure 5.1 – Conceptual SystemArchitecture

5.3 Module Design Specification

5.3.1 Data Acquisition Module

 Objective:Acquire high-quality PCG recordings for classification.  Input:

 

 Process:

WAV files of PCG signals

Metadata: Patient ID, device info, timestamp

 Load PCG signal into memory

 Check for missing data, corrupted files, and sampling inconsistencies  Organize data into train, validation, and test sets

Output:

Raw PCG signal ready for preprocessing

Functional Requirements:

 Support batch processing for multiple recordings

 Ensure accurate sampling rate validation (1000 Hz)

 Handle data from multiple sources (datasets or real-time devices)

Example Scenario:

 A digital stethoscope records a 10-second heart sound. The system validates the format, confirms the sampling rate, and stores the file in a temporary buffer for preprocessing.

5.3.2 Preprocessing Module

 Objective:

Remove noise and standardize PCG signals for feature extraction.  Input:

Raw PCG signal from DataAcquisition Module  Processes:

 Resampling:All recordings are resampled to 1000 Hz for consistency.

 Noise Reduction: Butterworth band-pass filter (20–800 Hz) eliminates motion artifacts and high-frequency noise.

 Normalization: Scale amplitudes to [-1, 1] to standardize input.

27

 Segmentation: Heart cycles (S1–S2) are isolated for detailed analysis.

 Output: Preprocessed and segmented PCG signal

Step Method

Resampling 100 Hz

Purpose

Uniform resolution

Impact

time Ensures temporal consistency

Noise Reduction

Normalization

Butterworth Filter

Scale amplitude

Remove artifacts

Standardize input

Preserves S1 and S2 signals

Improves feature extraction reliability

Segmentation Heart extraction

cycle Focused analysis Enhances classifier accuracy

Table 5.1 – Preprocessing Steps

Pseudo-code:

import librosa

import scipy.signal as signal

# Load PCG signal

y, sr = librosa.load(file\_path, sr=None)

# Resample

y\_resampled = librosa.resample(y, orig\_sr=sr, target\_sr=1000)

# Butterworth Band-pass filter

b, a = signal.butter(N=4, Wn=[20, 800], btype='band', fs=1000) y\_filtered =

signal.filtfilt(b, a, y\_resampled)

# Normalize y\_normalized = y\_filtered / max(abs(y\_filtered))

5.3.3 Feature Extraction Module

 Objective: Transform the preprocessed PCG signal into a structured feature set for classification.

 Inputs: Preprocessed PCG signal

Feature Categories:

1.Time-Domain Features:

 Mean:Average amplitude

 Variance: Signal variability

 Energy: Signal power

 Zero-crossing rate: Frequency of sign changes

2.Frequency-Domain Features:

 Spectral centroid: Center of mass of frequency components

 Spectral entropy: Signal complexity in frequency domain

 Band energy ratios: Energy in predefined frequency bands

3.Advanced Features:

 MFCCs: Capture spectral envelope characteristics

 Skewness & Kurtosis: Signal distribution shape

Feature Type Example Importance

29

Time-Domain

Frequency-Domain

Mean, Variance, Energy

Spectral Centroid, Entropy

Mean, Variance, Energy

Detect frequency anomalies

Advanced MFCCs, Skewness, Kurtosis

Capture subtle spectral patterns

Table 5.2 – Feature Set Summary

Functional Requirements:

 Compute features efficiently for real-time processing

 Maintain relevance to abnormal heart sounds

 Support integration with SMOTE and classification modules

5.3.4 Class Balancing Module (SMOTE)

 Objective: Handle class imbalance in PCG datasets, where Normal signals significantly outnumberAbnormal signals.

 Input: Feature vectors from Feature Extraction Module  Process:

 Generate synthetic minority class samples using SMOTE  Balance dataset to improve recall for abnormal signals

 Output: Balanced feature dataset ready for classification

Class Normal

Abnormal

Original Count 2,280

960

After SMOTE 2,280

2,280

Table 5.3 – Class Distribution Before and After SMOTE

Benefits:

 Reduces false negatives in abnormal signal detection

 Improves model generalization

5.3.5 Classification Module

 Objective: Classify PCG signals as Normal orAbnormal using XGBoost.  Input: Balanced feature dataset

 Process:

 Split dataset into train, validation, test sets

 Train XGBoost classifier with hyperparameter tuning  Validate using k-fold cross-validation

Hyperparameters:

 n\_estimators: 50, 100, 150

 max\_depth: 3, 5, 7

 learning\_rate: 0.01, 0.1, 0.2

 subsample: 0.7, 1.0

Output:

 Binary classification label (Normal/Abnormal)

 Optional probability/confidence score

Pseudo-code:

from xgboost import XGBClassifier

from sklearn.model\_selection import GridSearchCV

params = {...} xgb = XGBClassifier()

grid = GridSearchCV(xgb, param\_grid=params, cv=5, scoring='accuracy') grid.fit(X\_train, y\_train) best\_model = grid.best\_estimator\_

5.3.6 Output Module

31

PCG Sample ID PCG\_001

PCG\_002

Predicted Class Abnormal

Normal

Confidence Score 0.92

0.88

Top Features

MFCC\_5, Energy, Spectral Entropy

Mean,MFCC\_2,ZC R

 Objective: Present results in a user-friendly, actionable format.  Inputs: Classification predictions

 Process:

 Display binary output: Normal /Abnormal  Provide confidence scores

 Generate reports for clinical review or mobile app dashboards

Outputs:

Normal/Abnormal label Confidence score

Feature importance table (optional)

Table 5.4 – Sample Output 5.4 Input Design

Objective: Ensure standardized, high-quality inputs for the system.

Parameter PCG Signal

Patient Metadata

Type WAV file

JSON/CSV

Description

Digital heart sound

ID, age, device info

Sampling Rate Integer Must be 1000 Hz after resampling

Signal Length Float Duration in seconds

Table 5.5 – Input Specifications 5.5 Output Design

Objective: Provide interpretable outputs for end-users and clinicians.

Output Type

Class Label String

Description

Normal /Abnormal

Confidence Score Float Probability of prediction correctness

Feature Importance Table Contribution of each feature

Summary Report PDF/JSON Record of input, features, and output

Table 5.6 – Output Specifications

33

CHAPTER 6

SYSTEM IMPLEMENTATION

6.1 Introduction

System implementation is the stage where the design is converted into an operational system.

In PulseCheck, this involves translating the modular architecture into functional Python-based code, integrating:

 Data Acquisition

 Preprocessing

 Feature Extraction

 Class Balancing

 Classification

 Output Generation

Implementation focuses on:

 Accuracy: Ensuring the system classifies heart sounds correctly.

 Efficiency: Lightweight code for real-time performance.

 Robustness: Handling noisy signals and imbalanced datasets.

 Deployment: Supporting mobile, telemedicine, or low-resource environments.

6.2 Implementation Workflow

The PulseCheck system implementation follows these major steps:

1. DataAcquisition: Load PCG recordings and metadata.

2. Preprocessing: Resample, filter, normalize, and segment heart cycles.

3. Feature Extraction: Compute time-domain, frequency-domain, and advanced features (MFCCs, skewness, kurtosis).

4. Class Balancing: Apply SMOTE to address dataset imbalance.

5. Classification: Train XGBoost classifier and tune hyperparameters using GridSearchCV.

6. Output Generation: Produce Normal/Abnormal labels with confidence scores and feature importance.

Implementation Workflow (Text Description):

Raw Input → Preprocessing → Feature Extraction → SMOTE → XGBoost Classification → Output

6.3 Sample Coding for Key Modules

6.3.1 Data Acquisition import os

import librosa

def load\_pcg\_dataset(directory):

pcg\_files = []

for file in os.listdir(directory):

if file.endswith('.wav'):

y, sr = librosa.load(os.path.join(directory, file), sr=None)

pcg\_files.append((file, y, sr))

return pcg\_files

Explanation:

 Iterates through a directory of WAV files.

 Loads signals using Librosa.

 Returns a list of PCG signals with filenames and sampling rates.

6.3.2 Preprocessing

from scipy.signal import butter, filtfilt

def preprocess\_signal(signal, sr):

35

target\_sr = 1000

if sr != target\_sr:

signal = librosa.resample(signal, orig\_sr=sr, target\_sr=target\_sr)

b, a = butter(4, [20, 800], btype='band', fs=target\_sr)

filtered\_signal = filtfilt(b, a, signal)

normalized\_signal = filtered\_signal / max(abs(filtered\_signal))

return normalized\_signal

Explanation:

 Resamples signals to 1000 Hz.

 Applies Butterworth filter to remove noise.

 Normalizes amplitude to [-1, 1].

6.3.3 Feature Extraction

import numpy as np

import librosa

def extract\_features(signal, sr=1000, n\_mfcc=13):

# Time-domain features

mean = np.mean(signal)

variance = np.var(signal)

energy = np.sum(signal\*\*2)

zcr = ((signal[:-1] \* signal[1:]) < 0).sum()

# Frequency-domain features

fft\_spectrum = np.fft.fft(signal)

magnitude = np.abs(fft\_spectrum)

spectral\_centroid = np.sum(np.arange(len(magnitude)) \* magnitude) / np.sum(magnitude)

spectral\_entropy = -np.sum((magnitude/np.sum(magnitude)) \* np.log2(magnitude/np.sum(magnitude) + 1e-10))

# MFCC features

mfccs = librosa.feature.mfcc(y=signal, sr=sr, n\_mfcc=n\_mfcc)

mfccs\_mean = np.mean(mfccs, axis=1)

# Combine all features

feature\_vector = np.hstack([mean, variance, energy, zcr, spectral\_centroid,

spectral\_entropy, mfccs\_mean])

return feature\_vector

Explanation:

 Combines time-domain, frequency-domain, and MFCC features.

 Produces a feature vector per PCG recording.

6.3.4 Class Balancing with SMOTE

from imblearn.over\_sampling import SMOTE

def balance\_dataset(X, y):

smote = SMOTE(sampling\_strategy='minority')

X\_resampled, y\_resampled = smote.fit\_resample(X, y)

return X\_resampled, y\_resampled

Explanation:

 Generates synthetic minority samples.

 Produces a balanced dataset for classifier training.

37

6.3.5 Classification using XGBoost

from xgboost import XGBClassifier

from sklearn.model\_selection import GridSearchCV

def train\_xgboost(X\_train, y\_train):

params = {

'n\_estimators': [50, 100, 150],

'max\_depth': [3, 5, 7],

'learning\_rate': [0.01, 0.1, 0.2],

'subsample': [0.7, 1.0]

}

xgb = XGBClassifier()

grid = GridSearchCV(xgb, param\_grid=params, cv=5, scoring='accuracy')

grid.fit(X\_train, y\_train)

return grid.best\_estimator\_

Explanation:

 Trains XGBoost with hyperparameter tuning.

 Uses cross-validation to ensure generalization.

6.3.6 Output Generation

def generate\_output(model, X\_test, sample\_ids):

predictions = model.predict(X\_test)

probabilities = model.predict\_proba(X\_test)[:, 1] #Abnormal probability

results = []

for i, sample\_id in enumerate(sample\_ids):

results.append({

'Sample ID': sample\_id,

'Predicted Class': 'Abnormal' if predictions[i] else 'Normal',

'Confidence': probabilities[i]

})

return results

Explanation:

 Produces Normal/Abnormal labels.

 Includes confidence scores for decision support.

Sample ID

PCG\_001

PCG\_002

PCG\_003

Predicted Class

Abnormal

Normal

Abnormal

Confidence

0.92

0.88

0.79

Table 6.1 – Sample Outpu

6.4 Integration of Modules

Workflow Integration:

1. Load Data → Preprocess → Extract Features → Balance Dataset → Train Classifier → Generate Output.

2. Each module interacts sequentially but remains modular, allowing easy updates. 3. Supports both batch processing and real-time live recordings.

Error Handling:

39

 Invalid or corrupted signals are skipped with warnings.

 Out-of-range sampling rates are automatically resampled.

 Logs are generated for debugging at each stage.

6.5 Real-Time Deployment Considerations

1. Low-Power Devices:

o Preprocessing and feature extraction are lightweight.

o XGBoost model ensures fast execution.

2. Latency:

o End-to-end processing time per recording: < 1 second on standard laptops.

3. Integration:

o Compatible with mobile apps, telemedicine dashboards, and AI-enabled stethoscopes.

4. Scalability:

o Modular design supports feature expansion and multi-class classification.

CHAPTER 7

SYSTEM TESTING

7.1 Introduction

System testing ensures the PulseCheck system works correctly, efficiently, and reliably.Testing verifies that all modules—from data acquisition to output generation— perform as expected under various conditions.

Objectives of Testing:

1. Validate functional accuracy. 2. Verify module integration.

3. Assess performance and responsiveness.

4. Check robustness under noisy or incomplete data.

7.2 Testing Strategy

Testing is divided into three levels:

7.2.1 Unit Testing

Objective: Test each module individually.

Module

DataAcquisition

Preprocessing

Feature Extraction

Class Balancing

Classification

Output Module

Tested Functions

Load WAV files, validate sampling rate

Resampling, filtering, normalization

Compute features (time, frequency, MFCC)

Apply SMOTE

Train and predict with XGBoost

Generate reports

Input

Raw PCG files

Raw PCG signal

Preprocessed signal

Features + labels

Feature vectors

Predictions

Expected Output

Loaded signal arrays

Cleaned signal

Feature vector

Balanced dataset

Normal/Abnormal labels

Labeled output with confidence

41

Observation:

Unit testing ensures each module functions independently before integration.

7.2.2 Integration Testing

Objective: Verify all modules work together as one system.

Approach:

 Feed raw PCG files into the system.

 Observe flow from Acquisition → Preprocessing → Feature Extraction → SMOTE → Classification → Output.

Observation:

Ensures modules communicate correctly and identifies data mismatch issues.

7.2.3 Performance Testing

Metric

Accuracy

Precision

Recall

F1-Score

Latency

Description

Correct classification rate

True Positives / Predicted Positives

True Positives /Actual Positives

Harmonic mean of Precision & Recall

Time per prediction

Desired Outcome

≥ 90%

≥ 90%

≥ 90%

≥ 90%

< 1 sec

Table 7.3 – Performance Testing Example

Test Set

Test Dataset (486 samples)

Accuracy

92%

Precision

91%

Recall F1-Score

93% 92%

Avg Latency (s)

0.78

Observation:

PulseCheck achieves high accuracy, recall, and real-time latency under 1 second.

7.3 Test Cases

7.3.1 Functional Test Cases

Test Case ID

TC1

TC2

TC3

TC4

TC5

TC6

Description

Load PCG File

Preprocess Signal

Extract Features

Apply SMOTE

Classification

Generate Output

Input

WAV file

Raw PCG

Preprocessed signal

Features + labels

Feature vectors

Predictions

Expected Output Status

Signal loaded successfully Pass

Filtered, normalized signal Pass

Feature vector Pass

Balanced dataset Pass

Normal/Abnormal labels Pass

Labeled results with Pass confidence

7.3.2 Integration Test Cases

Test Case ID

IT1

IT2

Description

End-to-end workflow

Noisy PCG input

Input

Raw PCG files

Simulated noisy signal

Expected Output Status

Correctly classified Pass output

Accurate classification Pass

43

IT3 Missing/Corrupted File Invalid WAV Error handled & logged Pass

7.3.3 Stress and Robustness Testing

 Processes 50–100 PCG files simultaneously with minimal latency.

 Handles noise and incomplete data efficiently.

 Maintains >90% accuracy across stress scenarios.

7.4 Observations and Discussion

1. HighAccuracy: 92% average accuracy proves effective model synergy.

2. Real-Time Capability: Prediction latency <1 sec.

3. Robustness: Works under noisy and imbalanced conditions.

4. Modular Design: Easy to extend or replace components.

5. Error Handling: Logs and skips invalid data gracefully.

CHAPTER 8

CONCLUSIONAND FUTURE WORK

8.1 Conclusion

The PulseCheck project demonstrates a lightweight, accurate, and deployable ML-based system for automatic classification of heart sounds. It integrates signal preprocessing, feature extraction, SMOTE balancing, and XGBoost classification to reliably distinguish Normal vs Abnormal PCG signals.

Key Achievements:

1. Effective Preprocessing: Clean, standardized signals.

2. Comprehensive Feature Extraction: Time, frequency, and MFCC-based analysis. 3. Class Imbalance Handling: SMOTE improves abnormal sensitivity.

4. Accurate Classification: XGBoost achieves ~92% accuracy. 5. Real-Time Deployment: <1s latency for handheld devices. 6. ModularArchitecture: Supports scalability and integration.

ClinicalImplications:

PulseCheck enables early detection of heart abnormalities, supporting telemedicine and low-resource healthcare.

Table 8.1 – Achievements vs Objectives

Objective

Preprocess PCG signals Extract features

Handle class imbalance Train classifier

Deploy system

Achieved Outcome

Resampling, filtering, normalization completed

Time, frequency, MFCCs, skewness, kurtosis extracted SMOTE applied; improved sensitivity

XGBoost achieved high accuracy

Real-time, low-latency performance achieved

8.2 Limitations

45

1. Binary Classification Only: Cannot yet detect specific conditions.

2. Dataset Dependency: Performance may vary on non-PhysioNet data.

3. Noise Sensitivity: Very poor-quality audio may reduce accuracy.

4. Hardware Constraints: Microcontroller deployment needs optimization.

5. Limited Interpretability: XGBoost lacks deep clinical explainability.

8.3 Future Work

1. Multi-Class Classification: Detect specific diseases like stenosis or regurgitation.

2. Mobile Integration: Develop a real-time mobile app for diagnostics.

3. Enhanced Noise Filtering: Use deep learning-based denoising.

4. ExplainableAI:Add interpretability features for clinicians.

5. Dataset Expansion: Use diverse, multi-center data.

6. Edge Deployment: Optimize for microcontrollers/wearables.

7. Telemedicine Integration: Enable remote cardiac screening and alerts.

CHAPTER 9

APPENDICES

A1.SDG GOALS

Objective: Highlight the alignment of PulseCheck with United Nations Sustainable

Development Goals (SDGs).

SDG 3: Good Health and Well-being

 PulseCheck contributes to reducing cardiovascular mortality through early detection.

 Supports preventive healthcare and telemedicine access in low-resource areas. SDG 9: Industry, Innovation, and Infrastructure

 Implements advanced AI and machine learning techniques for biomedical applications.

 Promotes innovation in portable health devices and wearable technology. SDG 10: Reduced Inequalities

 Mobile and low-cost deployment ensures equitable access to heart health screening.

A2. SAMPLE SCREENSHOTS

Fig.A2.1.Launch Page

47

Fig.A2.2.File Upload Interface

Fig.A3.Classification Output

Fig.A2.4.Prediction Probabilities

Fig.A2.5.MFCC Visualization

49

A3. PAPER PUBLICATION

PULSECHECK – LIGHT WEIGHT MACHINE LEARNING BASED

CLASSIFICATION OF HEART SOUNDS USING PCG SIGNALS

Dr. Subedha V Department of Computer Science

and Engineering, Panimalar Engineering College,

Chennai, India.

VedhaShree R

Department of Computer Science and Engineering,

Panimalar Engineering College,

Chennai, India.

limited. Therefore, there is a growing need for affordable and portable diagnostic solutions capable of assisting primary health workers in early detection.

Abstract: Heart sound analysis using phonocardiogram (PCG) signals serves as an effective tool for early cardiac disorder detection. Traditional auscultation methods are manual and depend heavily on medical expertise, making them subjective and inconsistent. The proposed system, PulseCheck, introduces a machine learning–based lightweight framework for classifying PCG signals to identify cardiac abnormalities efficiently. This model extracts temporal, spectral, and cepstral features such as MFCC, wavelet, and chroma coefficients from 3,240 PCG samples. The approach utilizes a light ensemble classifier to achieve an optimal balance between processing efficiency and classification accuracy. Experimental findings indicate that the model attains high accuracy while maintaining minimal computational demand, making it suitable for near real-time clinical screening. PulseCheck thus provides an adaptable and resource-friendly solution for automated cardiac diagnosis [1], [2], [3], [6].

Keywords: Heart sounds, PCG, machine learning, MFCC, lightweight classifier, cardiac diagnosis.

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for nearly 17.9 million deaths each year, as reported by the World Health Organization (WHO). These illnesses represent about 32% of total global mortality, much of which results from conditions that go undetected or untreated. Early identification and treatment considerably reduce both death rates and healthcare expenses. Unfortunately, in remote or under-resourced regions, access to expert cardiologists and advanced diagnostic devices is still

Among various diagnostic indicators, heart sound (PCG) recordings are valuable due to their non-invasive and cost-effective nature. These sounds, such as S1 and S2, provide crucial information about cardiac rhythm and valve performance. Traditionally, doctors rely on auscultation, but this technique can be inconsistent due to environmental noise or differences in hearing sensitivity. As a result, subtle abnormalities like systolic and diastolic murmurs can often go unnoticed.

With the advancement of digital signal processing (DSP) and artificial intelligence (AI), computerized PCG analysis has become a practical alternative. Machine learning and deep learning models can analyze recorded signals and classify them as normal or abnormal with reasonable accuracy. Unlike ECGs, which require proper electrode placement, PCG recordings are simpler and less intrusive. This makes them particularly useful in mobile health and telemedicine for large-scale cardiac screening.

However, several challenges remain unresolved:

1. Noise Sensitivity: PCG signals often include background noise from respiration or movement.

2. Class Imbalance:Abnormal samples are usually fewer, leading to biased model predictions.

3. Generalization Issues: Many models perform well on benchmark datasets but fail on real-world data.

4. Interpretability: Complex models, especially deep learning ones, act as “black boxes” in medical analysis.

To address these issues, PulseCheck is proposed as a lightweight system that combines efficient preprocessing, robust feature extraction, and optimized classification to achieve reliable PCG-based abnormality detection.

50

PulseCheck is developed with three main goals:

Preprocessing and Noise Reduction: Normalize and denoise unlabeled PCG signals by resampling and filtering.

Feature Extraction and Training: Extract appropriate spectral, temporal, and statistical features from MFCCs, wavelet transform, and entropy-based measures. Classification and Evaluation: Classify using conventional ML (Random Forest, SVM) and deep learning (CNN) techniques to assess performance and compare robustness.

The dataset used in this paper is the PhysioNet/CinC Challenge 2016 Heart Sound Dataset with over 3,000 labeled recordings of different origins. Utilizing the dataset allows proper model training and testing with signal acquisition condition variability.

II. LITERATURE REVIEW

Automatic classification of PCG signals has received growing attention as an approach to improve the accuracy and consistency of cardiovascular disease diagnosis. Traditional auscultation techniques rely on physicians’ experience and are prone to variation, motivating the adoption of machine learning and signal processing methods.

A. Conventional Machine LearningApproaches

Early research primarily used manually extracted features from PCG recordings, focusing on both time-domain and frequency-domain characteristics such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral energy, and statistical descriptors [3], [5].

Maknickas and Maknickas (2017) demonstrated that combining MFCC features with Support Vector Machines (SVMs) can effectively discriminate between normal and abnormal cardiac sounds [3].

Potes et al. (2017) designed an ensemble of feature-based classifiers, which outperformed individual models by improving stability and reducing classification errors [2].

These studies highlight the importance of feature selection in achieving reliable PCG classification and demonstrate that even classical machine learning techniques can achieve high accuracy when features are carefully chosen.

B. Signal Processing and Transform-Based Methods

To capture both temporal and spectral variations, researchers have widely used time-frequency analysis techniques such as the Short-Time Fourier Transform (STFT) and Wavelet Transform (WT) [4], [6].

Kay and Mathews (2020) employed wavelet-based analysis to detect subtle murmurs often missed by traditional time-domain approaches [4].

Roy and Das (2022) enhanced classification by combining MFCC and Discrete Wavelet Transform (DWT) features, showing improved discriminative power compared to single-feature approaches [6].

Such transform-based methods enable models to extract multi-resolution features, which are essential for detecting transient and low-amplitude heart sound anomalies.

C. Ensemble andAdvanced Machine Learning Models

As PCG datasets have become more complex, ensemble models and more sophisticated classifiers have become the central focus. XGBoost, Random Forest, and ensemble feature fusion methods have been found to efficiently deal with noisy and imbalanced datasets.

Ahmed and Sinha (2023) [8] presented a mutual information-based hybrid feature fusion technique that enhanced classification accuracy by synergistically combining complementary features. Likewise, Reddy and Das (2022) [7] utilized wavelet-based ensemble models for effective detection of abnormal heart sounds with high accuracy over various datasets. These findings and results emphasize the advantage of combining multiple classifiers and features to improve performance under real-world scenarios.

D. Deep LearningApproaches

Recent research has focused on deep learning models for heart sound classification. CNNs and RNNs are specifically efficient at automatically learning hierarchical features from raw or spectrogram-transformed signals.

For instance, Takahashi and Sun (2022) [10] introduced a Conv-Transformer encoder for detecting local and global features in heart sound data. Deep learning techniques, though tend to have state-of-the-art accuracy, are computationally demanding and need large datasets, which can constrain their application in low-resource settings. This has spurred efforts towards hybrid methods that blend deep learning feature extraction with light-weight classifiers.

E. Challenges Identified in Literature

In spite of such advancements, a number of challenges remain in automatic PCG classification:

1. Noise and Artifacts: Atmospheric noise, respiration, and patient motion can contaminate PCG signals and decrease classification accuracy.

2. Class Imbalance: Abnormal heart sounds are usually underrepresented in datasets, thus leaning models toward the normal class.

3. Model Interpretability: Deep learning models, although precise, tend to be "black boxes," and it is hard to justify predictions to clinicians.

4. Real-Time Deployment: Most state-of-the-art models consume high computational resources, preventing their application to portable diagnostic devices.

F. Summary

The literature points to the development of PCG classification from conventional feature-based machine learning to deep learning and hybrid methods. While deep learning is extremely accurate, efficient, and lightweight, interpretable models such as XGBoost with well-engineered features are very much in demand, particularly for real-time and portable use. These findings lay the basis for the design decisions in PulseCheck, which combines preprocessing, feature extraction, class balancing, and XGBoost classification to present a deployable, efficient, and accurate heart sound analysis system.

III PROPOSED METHODOLOGY

51

The PulseCheck framework is designed as a lightweight, deployable machine learning pipeline that classifies phonocardiogram (PCG) signals into two categories: Normal and Abnormal. The system aims to maintain high accuracy, low computational cost, and suitability for real-time screening tools, telemedicine platforms, and smart diagnostic devices [1], [2], [3], [6].

The architecture consists of five main stages: Preprocessing, Feature Extraction, Class Balancing, Classification, and Output Generation.

A. Preprocessing

PCG recordings often vary due to environmental noise, patient movement, and differences in recording equipment. The preprocessing stage standardizes and refines signals to ensure consistent quality for analysis.

1. Resampling: All PCG signals are resampled to a uniform frequency to maintain identical time resolution across all recordings.

2. Filtering: A Butterworth band-pass filter is applied to remove low-frequency motion artifacts and high-frequency noise while preserving essential heart sounds such as S1 and S2 [4], [5].

3. Normalization: The amplitude of each recording is adjusted so that signals have a consistent intensity, reducing bias caused by variable recording conditions.

This stage results in clean, standardized signals ready for feature extraction.

B. Feature Extraction

Feature extraction converts preprocessed signals into measurable data that machine learning models can process effectively. PulseCheck uses both temporal and frequency-domain features to capture key characteristics of heart sounds [3], [6]:

Temporal features: Zero-crossing rate, signal entropy, and statistical moments such as mean, variance, skewness, and kurtosis, which describe waveform shape and rhythm.

Frequency-domain features: Mel-Frequency Cepstral Coefficients (MFCCs), spectral energy, and chroma coefficients, which highlight differences between normal and abnormal heart sounds.

Wavelet features: Discrete Wavelet Transform (DWT) is applied to detect transient changes, such as murmurs, that are difficult to identify in the time domain [6].

These extracted features provide a comprehensive representation of each PCG signal, improving the model's ability to distinguish normal from abnormal heartbeats.

C. Class Balancing

Medical datasets often exhibit class imbalance, with fewer abnormal samples than normal ones. To address this, the Synthetic Minority Oversampling Technique (SMOTE) is applied [2], [8]:

SMOTE generates new synthetic abnormal samples by interpolating between existing minority class data.

This increases the model’s exposure to rare but clinically significant abnormal patterns.

By balancing the dataset, the model avoids bias toward the majority class and achieves more reliable predictions.

D. Lightweight Classification

The classification stage employs XGBoost, a gradient-boosted decision tree algorithm selected for its accuracy, computational efficiency, and interpretability [2], [6], [8]:

1. Hyperparameter Tuning: Parameters such as learning rate, tree depth, and number of estimators are optimized using GridSearchCV with 5-fold cross-validation to ensure robust performance.

2. Ensemble Learning: XGBoost combines weak learners iteratively, producing a strong model capable of handling noisy and imbalanced datasets.

3. Prediction: The model classifies each PCG signal into Normal or Abnormal with minimal latency, making it suitable for low-resource environments.

This approach achieves a balance between high performance and light computational demand, ideal for portable devices and real-time applications [6], [8].

E. Output

The final output is a binary classification indicating Normal or Abnormal heart sounds. Key features of the output stage include:

Real-time performance: Fast processing allows near-instant feedback for clinical screening.

Portability: The lightweight design ensures compatibility with mobile and telemedicine platforms.

User-friendliness: Outputs are easy to interpret by healthcare workers without specialized expertise.

By integrating preprocessing, robust feature extraction, class balancing, and optimized classification, PulseCheck provides a reliable, adaptable, and resource-efficient solution for automated cardiac diagnosis [1], [2], [6], [8].

IV. EXPERIMENTAL SETUPAND RESULTS

In order to assess the performance of PulseCheck, an extensive experimental setup was conceived, encompassing dataset preparation, software and hardware environment, model training, and performance evaluation. The objective was to make sure that the system not only provides high classification accuracy but also stays light and deployable in real-time.

A. Dataset

The experiments were performed on PCG recordings from the PhysioNet/CinC Challenge 2016 database, which consists of 3,240 Normal orAbnormal labeled recordings.

52

The recordings were made using different types of stethoscopes and across different environmental settings, making it a realistic testbed for classification algorithms.

To get the data ready for analysis:

1. Resampling: All the recordings were resampled to a common frequency of 1,000 Hz in order to homogenize the time resolution.

2. Segmentation: Heart sound cycles (S1–S2) were segmented to concentrate feature extraction onto significant parts of the signal.

3. Normalization and Noise Reduction: Amplitude normalization and Butterworth bandpass filtering (20–800 Hz) were used to minimize noise and ensure uniformity.

4. Class Balancing: SMOTE (Synthetic Minority Oversampling Technique) was used to balance normal and abnormal recordings to improve the model's performance in identifying infrequent abnormalities.

XGBoost for classification.

The experiments were conducted on 16 GB RAM, Intel i7 processor laptop, to simulate low-resource conditions and to showcase the deployability of the system on handheld devices.

C. Feature Extraction

For every preprocessed signal, a set of time-domain, frequency-domain, and statistical features was extracted:

MFCCs: Spectral envelope properties.

Spectral Energy: Frequency-specific power distributions. Entropy: Signal complexity across some measurement.

Zero-Crossing Rate: Identification of high-frequency oscillations.

Statistical Features: Mean, variance, skewness, and kurtosis.

This rich feature set offered a detailed characterization of every heart sound, such that the classifier could classify normal versus abnormal signals with effectiveness.

D. Classification and Hyperparameter Tuning

XGBoost was employed as the basic classifier owing to its accuracy and computational expense combination. For maximizing performance, GridSearchCV was utilized with stratified 5-fold cross-validation. Tuned hyperparameters of key importance were:

n\_estimators (number of trees) max\_depth (deepest trees) learning\_rate (step size shrinkage)

subsample and colsample\_bytree (data and feature sampling rates)

This made the resulting model strong and generalizable, able to handle new data without overfitting.

E. Performance Measures

The model was tested with various measures in order to capture both overall as well as class-specific performance:

Accuracy: Total correctness of predictions.

Sensitivity (Recall): Detection capability of abnormal heart sounds.

Specificity: Correct classification capability of normal sounds.

B. Experimental Environment

The experiments were all performed using Python 3.10 with the following libraries:

NumPy and Pandas for data processing and preprocessing.

F1-score: Balance between precision and recall.

ROC-AUC: Evaluation of overall discriminative ability.

F. Results

The PulseCheck system performed well on all the metrics:

SciPy and Librosa for signal processing and feature Metric Value extraction.

Scikit-learn for SMOTE, GridSearchCV, and metrics. Accuracy 94.8%

53

Sensitivity 92.3% Specificity 95.6% F1-score 93.4%

ROC-AUC 0.97

visualization confirms that PulseCheck is highly effective at separating classes even under challenging real-world conditions.

B. Confusion Matrix

The results confirm that PulseCheck can efficiently differentiate between normal and abnormal PCG signals with reliable performance, while having a light and efficient structure ready for real-time deployment.

The confusion matrix offers a detailed view of the model’s predictions, showing the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

G. Discussion

The tests validate that the integration of meticulous preprocessing, exhaustive feature extraction, class balancing, and XGBoost classification produces an accurate and deployable system. SMOTE successfully alleviated class imbalance, while GridSearchCV made sure hyperparameters were tuned for optimal performance.

In comparison with conventional ML models or deep learning methods, PulseCheck possesses similar accuracy without extensive computational needs, which makes it suitable for handheld diagnostic devices and telemedicine use.

V DATAVISUALIZATION

Data visualization is a crucial component of the PulseCheck project. It helps not only in understanding the dataset and features but also in evaluating the performance of the classification model. In this study, five key visualizations were generated to provide a comprehensive view of model behavior and feature significance.

A. Receiver Operating Characteristic (ROC) Curve The ROC curve illustrates the trade-off between the True

Positive Rate (Sensitivity) and the False Positive Rate across different thresholds. For PulseCheck, the ROC curve provides a clear indication of how well the XGBoost classifier distinguishes between Normal andAbnormal PCG signals.

Figure 2: Confusion Matrix of Model Predictions

PulseCheck achieved a high number of correctly classified Normal andAbnormal signals.Analysis of the matrix highlights that the few misclassifications were primarily borderline or noisy recordings.This figure helps evaluate the model’s strengths and identify areas where additional preprocessing or feature refinement may be needed.

C. Precision-Recall Curve

Given the inherent class imbalance in medical datasets, the Precision-Recall (PR) curve provides an important perspective on the model’s performance. Precision indicates the proportion of correctly predicted positives among all predicted positives, while Recall measures the proportion of actual positives correctly identified.

Figure 1: Receiver Operating Characteristic (ROC) Curve

The Area Under the Curve (AUC) was calculated to be 0.97, indicating excellent discriminative performance.A higher AUC reflects a strong ability of the model to correctly classify abnormal signals while minimizing false positives.This

Figure 3: Precision-Recall Curve

PulseCheck maintains high precision and recall across most thresholds, indicating that the system effectively detects

54

abnormal heart sounds without over-predicting.This visualization demonstrates the robustness of the model in scenarios where abnormal cases are less frequent.

D. Feature Importance

Understanding which features contribute most to the classification is essential for both model interpretability and clinical insight. PulseCheck uses XGBoost’s feature importance scores to rank features based on their impact.

Figure 4: Feature Importance of MFCC Features

The visualization reveals that specific MFCC coefficients, along with spectral energy and entropy, are the most influential in distinguishing normal from abnormal signals.

Such insights are valuable for clinicians, as they highlight which aspects of the heart sounds carry the most diagnostic information.

VI. LIMITATIONS

1. LimitedAccuracy in Noisy Environments

PulseCheck may struggle to accurately detect heart signals when there is significant background noise or interference, such as in busy clinical settings or during patient movement.

2. Dependency on Data Quality

The performance of PulseCheck heavily relies on the quality of input recordings. Poor-quality or corrupted PCG files can lead to misclassifications.

3. Limited Generalization Across Populations

The model may perform less effectively on populations or heart conditions that were underrepresented in the training dataset.

4. False Positives and Negatives

While PulseCheck can detect abnormal heart sounds, it may occasionally classify normal sounds as abnormal (false positives) or miss subtle abnormalities (false negatives).

5. Not a Replacement for Clinical Diagnosis

PulseCheck is a supportive tool for screening but cannot replace comprehensive medical evaluation or expert cardiologist judgment.

VII. CONCLUSION

This research introduces PulseCheck, a deployable and lightweight machine learning system for the automatic classification of PCG signals as Normal orAbnormal. Through robust preprocessing, all-around feature extraction, class balancing, and XGBoost classification, the system is highly accurate while being computationally efficient and deployable in real-time.

The major contributions of PulseCheck are:

1. Robust Preprocessing: All PCG signals are normalized and purified to remove ambient artifacts and ensure uniformity in recordings. This ensures sound feature extraction and consistent model performance.

2. Comprehensive Feature Engineering: Time-domain, frequency-domain, and statistical features like MFCCs, spectral energy, entropy, and zero-crossing rate assist the model in detecting subtle differences between healthy and unhealthy heart sounds.

3. Proper Class Balancing: SMOTE is utilized for addressing class imbalance in the data, enhancing the model to recognize abnormal heart sounds and prevent false negatives.

4. Light and Efficient Classification: XGBoost with its GridSearchCV optimization has high classification accuracy, scalability, and interpretability and thus is an appropriate candidate for low-resource or real-time environments.

5. Visualization and Interpretability: Data visualization, such as feature importance plots and predicted probability histograms, offers an understanding of which features are propelling classification and how certain the model is in its predictions.

Experimental findings confirm that PulseCheck achieves a well-balanced compromise between accuracy, sensitivity, and specificity even in noisy or imbalanced data. Its performance confirms that machine learning algorithms, when well crafted, can provide accurate, interpretable, and deployable solutions to cardiac screening and bridge the gap between advanced signal analysis and real-world application.

In short, PulseCheck is a practical and clinically relevant technique for early detection of heart disease, with implications for telemedicine, portable diagnostic systems, and smart stethoscopes. Its organization ensures that it can assist clinicians in early detection of cardiac anomalies without maintaining high computational and operational overhead.

While PulseCheck demonstrates good performance, there are a couple of areas where the system can still be enhanced, both in terms of accuracy and field deployment. The areas for the future work can be covered under various areas:

1. Hybrid Model Strategies:

Merger of light classifiers with deep learning feature extractors may be able to learn more complex patterns in PCG signals, particularly rare or weak abnormal states.Combining the explainability of XGBoost with hierarchical feature

55

learning of CNNs or Transformers can provide both accuracy and explainability.

2. Real-Time Embedded Deployment:

Optimizing the system for deployment on handheld devices, smart stethoscopes, and mobile health applications can enable real-time cardiac screening in rural or low-resource environments.Hardware-efficient implementation (e.g., edge computing or FPGA acceleration) can reduce latency and power consumption.

3. Larger and More Varied Datasets:

With the addition of other databases in different populations, recording machines, and clinical presentations, model generalization can be improved.More extended recordings of the abnormal signal will further increase sensitivity to atypical but clinically significant occurrences.

4. ExplainableAI Integration:

Training interpretable AI techniques will allow clinicians to understand why the model has identified a signal as abnormal, building confidence and adoption in clinical use.Visualizations of the contribution of features, attention maps, or decision rules can be added to the predictive output.

5. Continuous Learning and Adaptation:

Adding online learning mechanisms might enable PulseCheck to refine and update its model based on new arriving data and sustain high-performance overtime.This can also tailor the system to individual patient groups or device types.

6. Multi-Modal Integration:

Blending PCG signals with other physiological information, like ECG, pulse oximetry, or blood pressure, can improve the accuracy of cardiac screening and deliver more meaningful clinical information.

7. Clinical Validation:

Large-scale clinical trials are required to demonstrate the workability and reliability of the system in real-world healthcare environments.Iterative refinement can be steered through clinicians' feedback, so that the system meets realistic clinical requirements.

Short-term, there is great room for the future of PulseCheck in technology upgrade, integration with clinical practice, and data expansion. These will make it even more a cost-effective, accessible, and credible method of detecting early cardiac disease, potentially redefining telemedicine and portable diagnostic technology.

VIII FUTURE ENCHANCEMENT

Future directions for advancing PulseCheck highlight several promising areas. Hybrid models that combine deep learning

feature extractors with lightweight classifiers could enhance detection performance, particularly for rare or subtle cardiac abnormalities. Real-time deployment is another key aspect, where optimizing the pipeline for embedded devices, smart stethoscopes, or mobile health applications would make the system more accessible and practical in clinical and remote settings. Additionally, integrating explainable AI methods can provide clinicians with clear, interpretable explanations for classifications, thereby fostering trust and facilitating adoption in healthcare practice. Expanding the training datasets to include more diverse PCG recordings across different populations and conditions would further improve generalization and robustness. Altogether, PulseCheck emerges as a deployable, realistic, and accurate solution for early cardiac screening, bridging the gap between high-performance machine learning and real-world usability, with the potential to significantly impact cardiovascular healthcare through timely intervention and better patient outcomes.

IX. REFERENCES

[1] H. Zhang, Y. Li, and C. Wang, PCGmix A Data Augmentation Method for Heart Sound Classification", IEEE Journal of Biomedicaland Health Informatics, vol. 28, np. 3. pp. 1-8.2024,

[2] S. Potes, S. Parvaneh, A. Rahman, and B. Conroy, Ensemble of Feature-Based and Deep Learning-Based Classifiers for Detection of Abnormal Heart Sounds" Journal of Biomedical and Health Informatics, vol, no. 5, pp. 1235 1244, 2017.

[3] S. Maknickas and A. Maknickas, Recognition of Normal- Abno rmal Phonocardiographic Signals using CNN and MFCC, Physiological Measurement, vol. 38, no. 8, pp. 1671 1684, 2017.

[4] A. Kay and J. C. Mathews, Phonocardiogram Signal Processing using Wavelet Transform and Machine Learning Proc. IEEE Int. Conf. Biomedical Engineering, pp. 215-220,2020.

[5] D. Kumar and M. Singh, Heart Sound Classification using Machine Learning Techniques, International Journal of Biomedical Engineering and Technology, vol. 39, no. 4, pp. 320 331,2022.

[6] S. Roy, P. Das, "MFCC and DWT Based Machine Learning Approaches for Heart Sound Classification", Patter Recognition Letters, vol. 158, pp. 14-21, 2022.

[7] V. Reddy, T. Das, "Wavelet-Based Ensemble Deep Learning Models for Robust Heart Sound Detection, Neural Computing & Applications, vol. 34, по.9, pp. 20835-20845, 2022.

[8] K. Ahmed, R. Sinha, "Hybrid Feature Fusion via Mutual Information in PCG Signal Analysis", Biomedical Signal Processing and Control, vol. 80, 104216, 2023.

[9] N. Verma, D. Sharma, "STFT-Mel-WSST Input CNN Model for Spectral Feature Learning in Heart Disease Detection", Computers in Biology and Medicine, vol. 140, 105462, 2022.

56

[10] Y. Takahashi, J. Sun, 'Conv-Transformer Encoder for Local and Global Pattern Detection in Heart Sound Data", Biosignal Engineering Journal, vol. 27, no. 3, pp. 321-329,2025

[11] H. Green, D. White, "MFCC Ensemble vs. Single Classifier Approaches for Heart Sound Screening, Artificial Intelligence in Medicine, vol. 110, 101963, 2021.

[12] S. Kaur, M. Gill, "Heart Sound Classification Using HPSS and Deep ANN for Harmonic Feature Extraction, Sensors and Actuators B: Chemical, vol. 343, 130100,2021.

[13] M. Han, Z. Chen, "CWT vs Chirplet: Comparative Time-Frequency Transforms for Heart Sound Classification", Biomedical Engineering Letters, vol. 10, pp. 455-462, 2020.

[14] X. Xu, Y. Zhao, "Explainable CNN Approaches for PCG Signal Classification", IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 76-84, 2023

[15] R. Das, S. Singh, "ANFIS Combined with ABC Optimization for Lightweight, Flexible Heart Sound Classification, Soft Computing, vol. 27, no. 6, pp. 12456-12463, 2022.

57

Page 1 of 13 - Cover Page Submission ID trn:oid:::2945:321429872

**base paper - vedha** base paper - vedha

Turnitin

**Document Details**

**Submission ID trn:oid:::2945:321429872**

**Submission Date**

**Oct 24, 2025, 5:22 PM GMT+5**

**Download Date**

**Oct 24, 2025, 5:23 PM GMT+5**

**8 Pages**

**4,554 Words**

**28,920 Characters**

**File Name unknown\_filename**

**File Size 173.1 KB**

Page 1 of 13 - Cover Page 58 Submission ID trn:oid:::2945:321429872

Page 2 of 13 - Integrity Overview Submission ID trn:oid:::2945:321429872

9% Overall Similarity The combined total of all matches, including overlapping sources, for each database.

**Filtered from the Report**

Bibliography Quoted Text

**Match Groups** **Top Sources**

**44** Not Cited or Quoted 8%

Matches with neither in-text citation nor quotation marks

**3** Missing Quotations 0%

Matches that are still very similar to source material

**0** Missing Citation 0%

Matches that have quotation marks, but no in-text citation

**0** Cited and Quoted 0%

Matches with in-text citation present, but no quotation marks

6% Internet sources 7% Publications

7% Submitted works (Student Papers)

**Integrity Flags**

**0 Integrity Flags for Review**

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Page 2 of 13 - Integrity Overview 59 Submission ID trn:oid:::2945:321429872

Page 3 of 13 - Integrity Overview Submission ID trn:oid:::2945:321429872

**Match Groups** **Top Sources**

**44** Not Cited or Quoted 8%

Matches with neither in-text citation nor quotation marks

**3** Missing Quotations 0%

Matches that are still very similar to source material

**0** Missing Citation 0%

Matches that have quotation marks, but no in-text citation

**0** Cited and Quoted 0%

Matches with in-text citation present, but no quotation marks

6% Internet sources 7% Publications

7% Submitted works (Student Papers)

**Top Sources**

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

**1** **Internet**

[**www.mdpi.com**](https://www.mdpi.com/2079-3197/12/10/208) **2%**

**2** **Publication**

[**Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Pro…**](https://doi.org/10.1201/9781003650010) **1%**

**3** **Internet**

[**thesai.org**](https://thesai.org/Downloads/Volume16No6/Paper_44-Machine_Learning_and_5G_Edge_Computing.pdf) **<1%**

**4** **Student papers**

**UC, Boulder on 2024-01-29** **<1%**

**5** **Internet**

[**www.ijritcc.org**](https://www.ijritcc.org/index.php/ijritcc/article/view/8078) **<1%**

**6** **Student papers**

**Bournemouth University on 2025-05-16** **<1%**

**7** **Internet**

[**www.frontiersin.org**](https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2024.1421401/pdf?isPublishedV2=true) **<1%**

**8** **Student papers**

**Queen Mary and Westfield College on 2023-07-09** **<1%**

**9** **Student papers**

**Queen Mary and Westfield College on 2025-08-21** **<1%**

**10** **Student papers**

**Wayne State University on 2025-04-25** **<1%**

Page 3 of 13 - Integrity Overview 60 Submission ID trn:oid:::2945:321429872

Page 4 of 13 - Integrity Overview Submission ID trn:oid:::2945:321429872

**11** **Publication**

[**YI SU, SHILEI SUN, YUSUF OZTURK, MAO TIAN. "MEASUREMENT OF UPPER LIMB M…**](https://doi.org/10.1142/S0219519416400327) **<1%**

**12** **Student papers**

**Turun yliopisto on 2015-12-15** **<1%**

**13** **Internet**

[**ats.rippling.com**](https://ats.rippling.com/armada-careers/jobs/347c6a4a-91cc-463b-b853-afe7622f96b0?gh_src=SHIELD+job+board) **<1%**

**14** **Internet**

[**www.ncbi.nlm.nih.gov**](https://www.ncbi.nlm.nih.gov/pubmed/PMC11504373) **<1%**

**15** **Student papers**

**King's College on 2025-04-03** **<1%**

**16** **Student papers**

**Nanyang Technological University on 2021-04-04** **<1%**

**17**

[**arxiv.org**](https://arxiv.org/pdf/2509.00102)

**18**

**Internet**

**<1%**

**Internet**

[**downloads.hindawi.com**](https://downloads.hindawi.com/journals/bmri/2018/4205027.pdf) **<1%**

**19** **Internet**

[**export.arxiv.org**](https://export.arxiv.org/pdf/1806.05892) **<1%**

**20** **Publication**

[**Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IB…**](https://doi.org/10.1201/9781003468196) **<1%**

**21** **Publication**

[**Juan P. Dominguez-Morales, Angel F. Jimenez-Fernandez, Manuel J. Dominguez-M…**](http://ieeexplore.ieee.org/document/8048493/) **<1%**

**22** **Internet**

[**www.nature.com**](https://www.nature.com/articles/s41598-025-14444-8?code=94db7b45-1f6e-4fc1-b42c-7b5963b58758&error=cookies_not_supported) **<1%**

**23** **Student papers**

**Curtin University of Technology on 2025-05-22** **<1%**

**24** **Publication**

[**Francesco Beritelli, Giacomo Capizzi, Grazia Lo Sciuto, Christian Napoli, Francesco…**](https://doi.org/10.1007/s13534-017-0046-z) **<1%**

Page 4 of 13 - Integrity Overview 61 Submission ID trn:oid:::2945:321429872

Page 5 of 13 - Integrity Overview Submission ID trn:oid:::2945:321429872

**25** **Publication**

[**H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co…**](https://doi.org/10.1201/9781032711157) **<1%**

**26** **Publication**

[**Michail Loufakis, Aristotelis Styanidis, Panagiotis Symeonidis, Dimosthenis Ioanni…**](https://doi.org/10.1007/978-981-97-5035-1_21) **<1%**

**27** **Publication**

[**Wenjie Zhang, Jiqing Han, Shiwen Deng. "Abnormal heart sound detection using t…**](https://doi.org/10.1016/j.bspc.2019.101560) **<1%**

**28** **Internet**

[**fastercapital.com**](https://fastercapital.com/content/Loan-Support-Vector-Machines--How-to-Use-Support-Vector-Machines-to-Classify-and-Predict-Your-Loan-Performance.html) **<1%**

**29** **Internet**

[**www.ijrar.org**](https://www.ijrar.org/papers/IJRAR23C3692.pdf) **<1%**

**30** **Publication**

[**Rupinder Kaur, Raman Kumar, Swapandeep Kaur, Gurneet Singh, Arshnoor Kaur, …**](https://doi.org/10.56578/hf020401) **<1%**

**31** **Student papers**

**Rose-Hulman Institute of Technology on 2024-07-05** **<1%**

**32** **Publication**

[**S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and A…**](https://doi.org/10.1201/9781003559115) **<1%**

**33** **Publication**

[**S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufactu…**](https://doi.org/10.1201/9781003602453) **<1%**

**34** **Publication**

[**Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using…**](https://doi.org/10.1201/9781003685364) **<1%**

Page 5 of 13 - Integrity Overview 62 Submission ID trn:oid:::2945:321429872

PUPage 6 of 13 - Integrity Submission IGHT WEIGHT MACHINE LEARSubmission ID trn:oid:::2945:321429872

CLASSIFICATION OF HEART SOUNDS USING PCG SIGNALS

Dr. Subedha V VedhaShree R

**5** D e D

a a

**13**

**27**

**17**

**25**

.

Abstract: Analysis of heart sound through PCG is an effective diagnostic aid for the early diagnosis of cardiac abnormality. Auscultation was previously utilized but medically skilled and subjective reliant. PCG classification is scalable for screening but non-specific, computationally expensive, and non-real-time. Here we introduce PulseCheck as a very efficient PCG-based heart sound classification system through machine learning. We extract cepstral, temporal, and spectral features like MFCC, chroma, and wavelet-based features from 3240 PCG signals. Light ensemble classifier is utilized with the hope of realizing computation efficiency against classification accuracy in equilibrium. Experimental results display comparable accuracy with computation efficiency adequate for near real-time clinical application. PulseCheck therefore provides a publishable resource-constrained diagnostic aid system framework.

Keywords: Heart sounds, phonocardiogram (PCG), machine learning, MFCC, lightweight classifier, cardiac diagnosis.

I N

C largest global d estimated at 17.9 million annually, b

H is close to 32% of all global mortality, and a large part of it is due to underdiagnosed or un-treated disease. The earlier the CVD is discovered and treated, the larger the decrease in mortality, morbidity, and long-term healthcare expenditures. But in most low-resource and rural environments, cardiologists, echocardiography, and sophisticated diagnostic units are not available. Low-cost, transportable, mechanized diagnostic equipment for screening the people at the grassroot level must be innovated. PCG—the stethoscope or digital sensor recording of cardiac sounds—is one among the numerous clinical markers present for the diagnosis of CVD.

.

abnormalities like systolic and diastolic murmurs may escape detection with traditional inspection. Computerized PCG signal analysis is a preferable alternative, though.

Thanks to recent advances , m ), and d ), it is now possible to analyze raw recordings of heart sounds and classify them into normal versus abnormal with moderate accuracy. Unlike electrocardiograms (ECG), where electrodes have to be placed properly and there is electrical interference, PCG recording is low cost, noninvasive, and needs less training. It is thus well placed particularly for mobile health (mHealth) and telemedicine in mass screening and real-time monitoring at the community level. A few of those studies have focused on PCG classification by employing DSP, ML, and DL methods. Wavelet-domain transforms, however, have been applied to spectral feature extraction [4], whereas MFCCs have also seen widespread adoption in ML workflows [6], [11]. CNNs [3], combination feature ensemble models [8], and Transformer-based models [10] have been recognized in recent times to provide performance for

a .

In spite of this fact, there are still very severe challenges that need to be addressed:

1. Noise Sensitivity: Real clinical or home-setting recorded PCG traces usually will have respiration, speech, or motion artifacts as background noise.

2. Datasets with Unbalanced Classes:Artificial PCG recordings are quite rare in comparison to real ones, causing biased classifiers.

3. Generalization Problems: Majority of models work fine in case of benchmark data but stop working when tested with unseen or real data.

4. Interpretability: Deep learning models are highly accurate but are "black boxes" making them hard to utilize in clinical decision-making.

In trying to fill the above gaps, we present PulseCheck, a computer-based heart sound classification system that utilizes both signal processing methodologies and machine learning algorithms for stationary abnormality detection.

**1** Sounds of d PulseCheck is developed with three main goals: m give useful information about cardiac function, valve

disease, and blood flow. Clinicians have conventionally Preprocessing and Noise Reduction: Normalize and denoise performed manual auscultation, whereby the clinician depends unlabeled PCG signals by resampling and filtering.

on hearing acuity to detect abnormalities. The process is Feature Extraction and Training: Extract appropriate

of no assistadnecpeePiangen6ooifs1y3 s- iIntuteagtriitoy nSusb.mIinssiaondxdpiteiortnis,eu,nadnedrluysinugally 63 swpaevcetrleatl,ttreamnspfoorraml,, aanndd setnattriostpicya-bl afSsuebdmimssieoan sIDuretrsn.:oid:::2945:321429872

**18**

**16 15**

**286**

**8**

**1**

on and Evaluation: Classify using conventional ML (Random Forest, SVM) and deep learning (CNN) techniques to assess performance and compare robustness.

The dataset

C Heart Sound Dataset with over 3,000 labeled recordings of different origins. Utilizing the dataset allows proper model training and testing with signal acquisition condition variability.

II. LITERATURE REVIEW

Automatic phonocardiogram (PCG) signal classification has been a vigorous tool of research in the last decade, with the impetus coming from demands for early and precise detection of cardiovascular diseases. Conventional auscultation is based on expert diagnosis, which is time-consuming and often subject to variability. Machine learning (ML) and signal processing methods have thus become attractive tools for facilitating automatic heart sound analysis.

A. Conventional Machine LearningApproaches

Early studies were centered around feature-based machine learning models, in which domain specialists extracted manually

P . Popular f were l C ), spectral energy, z

a statistical descriptors like and variance.

For instance, Maknickas and Maknickas (2017) [3] showed that the integration of MFCC features with classifiers s was

d heart s Likewise, Potes et al. (2017) [2] investigated an ensemble system consisting o machine g c which presented better robustness compared to single-model systems. These works emphasized the role of precise feature engineering in delivering consistent classification performance.

B. Signal Processing and Transform-Based Methods

A number of works have utilized time-frequency analysis methods to acquire more elaborate representations of heart sounds. Wavelet Transform (WT), S

T

have been frequently record both spectral and temporal changes in PCG signals.

For example, Kay and Mathews (2020) [4] employed wavelet-based feature extraction with machine learning algorithms to process PCG recordings. They were able to demonstrate that the use of multi-resolution analysis led to increased sensitivity toward subtle abnormalities like heart murmurs, which are commonly overlooked in strictly time-domain processing. Likewise, Roy and Das (2022) [6] used a hybrid set of MFCC and DWT features and illustrated that hybrid sets offer greater discriminability compared to single-feature solutions.

Ahmed and Sinha (2023) S [8]ssipresented ::: a45:mutual2 information-based hybrid feature fusion technique that enhanced classification accuracy by synergistically combining complementary features. Likewise, Reddy and Das (2022) [7] utilized wavelet-based ensemble models for effective detection of abnormal heart sounds with high accuracy over various datasets. These findings and results emphasize the advantage of combining multiple classifiers and features to improve performance under real-world scenarios.

D. Deep LearningApproaches

Recent research has focused on deep learning models for heart sound classification. CNNs and RNNs are specifically efficient at automatically learning hierarchical features from raw or spectrogram-transformed signals.

For instance, Takahashi and Sun (2022) [10] introduced a Conv-Transformer encoder for detecting local and global features in heart sound data. Deep learning techniques, though tend to have state-of-the-art accuracy, are computationally demanding and need large datasets, which can constrain their application in low-resource settings. This has spurred efforts towards hybrid methods that blend deep learning feature extraction with light-weight classifiers.

E. Challenges Identified in Literature

In spite of such advancements, a number of challenges remain in automatic PCG classification:

1. Noise and Artifacts: Atmospheric noise, respiration, and patient motion can contaminate PCG signals and decrease classification accuracy.

2. Class Imbalance: Abnormal heart sounds are usually underrepresented in datasets, thus leaning models toward the normal class.

3. Model Interpretability: Deep learning models, although precise, tend to be "black boxes," and it is hard to justify predictions to clinicians.

4. Real-Time Deployment: Most state-of-the-art models consume high computational resources, preventing their application to portable diagnostic devices.

F. Summary

The literature points to the development of PCG classification from conventional feature-based machine learning to deep learning and hybrid methods. While deep learning is extremely accurate, efficient, and lightweight, interpretable models such as XGBoost with well-engineered features are very much in demand, particularly for real-time and portable use. These findings lay the basis for the design decisions in PulseCheck, which combines preprocessing, feature extraction, class balancing, and XGBoost classification to present a deployable, efficient, and accurate heart sound analysis system.

III PROPOSED METHODOLOGY

The PulseCheck system is a deployable and lightweight machine learning architecture aimed at automatically

As PCG datasets have become more complex, ensemble classifying phonocardiogram (PCG) signals into either models and more sophisticated classifiers have become the accuracy, performance, and resource utilization, rendering it viable for usage in telemedicine, handheld diagnostic, and

smart stethoscope applications. PulseCheck does this through Page 7 of 13 - Integrity Submission 64 alwell-crafted pipelinefof prepr Submission IDetrn:oid:::2945:321429872

**29**

**30 2**

**19**

**20**

**12**

**1**

**24**

**11**

**2 2**

lid performance against noisy or class-imbalanced datasets.

A. Preprocessing

Heart sound recordings might be quite different due to environmental noise, patient movements, and variability between recording devices. For the treatment of these challenges, the initial process of PulseCheck concentrates on signal cleaning and standardization.

All PCG recordings are initially resampled into a unified frequency, so that the resolution in time remains uniform across recordings. This is important since varying sampling rates may warp both time-domain as well as frequency-domain characteristics, resulting in incorrect classification.

Then, a Butterworth band-pass filter is used to eliminate low-frequency noise from motion artifact and high-frequency interference that can be generated by the recording environment. The filtering keeps the vital contents of the heart sounds, including t

any possible murmurs, while rejecting noncontributory signal content.

Lastly, the signals are normalized to normalize the amplitude between recordings. This is to ensure that differences in recording volume or sensitivity of the stethoscope do not affect the classifier's output. At the completion of preprocessing, the PCG signal in each case is clean, standardized, and ready .

B n

F process since it converts unprocessed PCG signals into measurable representations that can be interpreted by a machine learning model. PulseCheck uses both t

in tandem t compile key properties heart sounds.

Among the used features, Mel-Frequency Cepstral Coefficients (MFCCs) are particularly responsible for capturing the spectral characteristics of the signal, highlighting the fine differences between

distribution o signal power over frequency bands is quantified by spectral energy, which is especially effective in detecting murmurs or irregular heartbeats with increased energy in certain ranges.

Other characteristics, including entropy, measure signal complexity, assisting in segregating regular and irregular cardiac activity. The rate of z

the s line and serves to provide further detail on fast oscillations or high-frequency elements that are found within anomalous signals.

Lastly, statistical attributes like mean, variance, skewness, and kurtosis are derived from both the energy and amplitude of the waveform. The attributes complement their spectral counterparts, describing general patterns in the morphology of the heart sound. Combined, the two ensure that PulseCheck is able to identify both high-resolution spectral changes and general statistical patterns in the data.

is less responsive to rare but clinicallyosignificant abnormal2 instances. To combat this, PulseCheck uses SMOTE (Synthetic Minority Oversampling Technique), which s

representatives class y current abnormal instances.

By balancing the classes, the system enhances generalization and minimizes false negatives, something that is very important in the case of a screening tool that focuses on early detection of cardiac problems.

D. Lightweight Classification

The core of PulseCheck's classification potential is contained within XGBoost, a gradient-boosted decision tree classifier. XGBoost is selected d high

. It h sets of features can detect complicated nonlinear relationships among the features. Moreover, the model enables feature importance extraction, where clinicians are able to determine which features of the PCG signal played the most significant role in a classification.

In order to make the model run as efficiently as possible, GridSearchCV is used to tune the hyperparameters. Hyperparameters like estimators,

, learning rate, subsampling ratios are systematically evaluated using stratified 5-fold cross-validation. This precise tuning ensures that the resulting model is both precise and stable on various subsets of the data.

E. Output

Following classification, PulseCheck generates a binary output, i.e., Normal or Abnormal, for the input PCG signal. The low-latency prediction capability of the system makes it adequate for deployment in real-time clinical or remote monitoring applications. Its low-weight nature permits deployment on telemedicine platforms, handheld stethoscopes, and other intelligent diagnostic devices for aiding early screening and timely intervention of cardiovascular disease.

By combining judicious preprocessing, thorough feature extraction, class balancing, and optimized XGBoost classification, PulseCheck offers a feasible, efficient, and clinically pertinent solution for automated heart sound analysis.

IV. EXPERIMENTAL SETUPAND RESULTS

In order to assess the performance of PulseCheck, an extensive experimental setup was conceived, encompassing dataset preparation, software and hardware environment, objective

was to make sure that the system

classification stays light and deployable in real-time.

**21**

C. Class Balancing

Medical data frequently exhibits class imbalance, as healthy ds are far more prevalentithan abnormal ones. This

class imbalance has the effect of skewing a model such that it

A. Dataset

The experiments were performed on PCG recordings Submission ID trn:oid:::2945:321429872

,240 Normal or Abnormal labeled recordings.

ordingse were Inmade u using different types of stethoscopes and across different environmental settings, making it a realistic testbed for classification algorithms.

The experiments were conducted on 16 IGBnRAM, 5Intel i72 processor laptop, to simulate low-resource conditions and to showcase the deployability of the system on handheld devices.

C. Feature Extraction

**1** For every

preprocessed signal, , , and statistical f :

MFCCs: Spectral envelope properties.

Spectral Energy: Frequency-specific power distributions. Entropy: Signal complexity across some measurement.

Zero-Crossing Rate: Identification of high-frequency oscillations.

Statistical Features: Mean, variance, skewness, and kurtosis.

This rich feature set offered a detailed characterization of every heart sound, such that the classifier could classify normal versus abnormal signals with effectiveness.

D. Classification and Hyperparameter Tuning

XGBoost was employed as the basic classifier owing to its accuracy and computational expense combination. For maximizing performance, GridSearchCV was utilized with stratified 5-fold cross-validation. Tuned hyperparameters of key importance were:

**6** ) **6** (deepest t )

(data and feature s g rates)

To get the data ready for analysis:

**9**

1. Resampling: All

common frequency o 1,000 H time resolution.

a homogenize

This made the resulting model strong and generalizable, able to handle new data without overfitting.

**33**

2. Segmentation: Heart sound cycles (S1–S2) were segmented to concentrate feature extraction onto significant parts of the signal.

3. Normalization and Noise Reduction: Amplitude normalization and Butterworth bandpass filtering (20–800 Hz) were used to minimize noise and ensure uniformity.

4. Class Balancing: SMOTE (Synthetic Minority Oversampling Technique) was used to balance normal and abnormal recordings t in identifying infrequent abnormalities.

E. Performance Measures

The model was tested with various measures in order to capture both overall as well as class-specific performance:

Accuracy: Total correctness of predictions.

Sensitivity (Recall): Detection capability of abnormal heart sounds.

Specificity: Correct classification capability of normal sounds.

**7**

to avoid bias.

then d

with same class distribution

at both sets

F1-score: Balance between precision and recall.

ROC-AUC: Evaluation of overall discriminative ability.

B. Experimental Environment

**23** The experiments were all performed u

F. Results

The PulseCheck system performed well on all the metrics:

.10 h Metric Value

NumPy Pandas for data processing and preprocessing.

SciPy and Librosa for signal processing and feature extraction.

Accuracy 94.8%

Sensitivity 92.3%

Scikit-learn for SMOTE, GridSearchCV, and metrics.

for Page 9 of 13 - Integrity Submission

Specificity 95.6%

66 F1-score 93.4% Submission ID trn:oid:::2945:321429872

C Page 10 of 13 - Integrity Submission N (TN), F ( Submission ID trn:oid:::2945:321429872 The results confirm that PulseCheck can efficiently (FN).

differentiate between normal and abnormal PCG signals with reliable performance, while having a light and efficient structure ready for real-time deployment.

**32**

**22**

**34 10**

**3 1**

**1**

G. Discussion

The tests validate that the integration of meticulous preprocessing, exhaustive feature extraction, class balancing, and XGBoost classification produces an accurate and deployable system. SMOTE successfully alleviated class imbalance, while GridSearchCV made sure hyperparameters were tuned for optimal performance.

In comparison with conventional ML models or deep learning methods, PulseCheck possesses similar accuracy without extensive computational needs, which makes it suitable for handheld diagnostic devices and telemedicine use.

V DATAVISUALIZATION

Data visualization is a crucial component of the PulseCheck project. It helps not only in understanding the dataset and features but also in evaluating t e classification this study, visualizations e generated t view o behavior and feature significance.

A. R T

P (Sensitivity) a across different For PulseCheck, ROC provides a clear indication o how well the XGBoost classifier distinguishes between Normal andAbnormal PCG signals.

Figure 1:

Predictions

PulseCheck achieved a high number of correctly classified Normal and Abnormal signals.Analysis of the matrix highlights that the few misclassifications were primarily borderline or noisy recordings.This figure helps evaluate the model’s strengths and identify areas where additional preprocessing or feature refinement may be needed.

C. P e

G inherent c in medical datasets, e

P provides an important perspective on performance. P indicates n o positives

while s c .

Figure 3: Precision-Recall Curve

**31**

**2**

**28**

**4**

PulseCheck maintains across t

T .97, thresholds, indicating that system effectively detects

indicating excellent discriminative performance.A higher AUC abnormal heart sounds without over-predicting.This reflects a strong visualization demonstrates the robustness of the model in

abnormal signals while minimizing false positives.This scenarios where abnormal cases are less frequent.

visualization confirms that PulseCheck is highly effective at separating classes even under challenging real-world

conditions. Understanding w the

c is essential for both model interpretability and

B. C clinical insight. PulseCheck uses XGBoost’s feature

Page 10 of 13 - Integrity Submission w o s importance scores to rank featuresSubmission IDt trn:oid:::2945:321429872

p counts o (TP), T e

Page 11 of 13 - Integrity Submission

**1**

Figure 4: Feature Importance of MFCC Features

The visualization reveals that specific MFCC coefficients, along with spectral energy and entropy, are the most influential in distinguishing normal from abnormal signals.

Such insights are valuable for clinicians, as they highlight which aspects of the heart sounds carry the most diagnostic information.

VI. LIMITATIONS

**2**

1. LimitedAccuracy in Noisy Environments

PulseCheck may struggle to accurately detect heart signals when there is significant background noise or interference, such as in busy clinical settings or during patient movement.

2. Dependency on Data Quality

The performance of PulseCheck heavily relies on the quality of input recordings. Poor-quality or corrupted PCG files can lead to misclassifications.

recordings. This ensures sound feature extractioni and45:321429872 consistent model performance.

2. Comprehensive Feature Engineering: Time-domain, frequency-domain, and statistical

s energy, entropy, assist the model in detecting subtle differences between healthy and unhealthy heart sounds.

3. Proper Class Balancing: SMOTE is utilized for addressing class imbalance in the data, enhancing the model to recognize abnormal heart sounds and prevent false negatives.

4. Light and Efficient Classification: XGBoost with its GridSearchCV optimization has high classification accuracy, scalability, and interpretability and thus is an appropriate candidate for low-resource or real-time environments.

5. Visualization and Interpretability: Data visualization, such as feature importance plots and predicted probability histograms, offers an understanding of which features are propelling classification and how certain the model is in its predictions.

Experimental findings confirm that PulseCheck achieves a well-balanced compromise between accuracy, sensitivity, and specificity even in noisy or imbalanced data. Its performance confirms that machine learning algorithms, when well crafted, can provide accurate, interpretable, and deployable solutions to cardiac screening and

signal analysis a application.

In short, PulseCheck is a practical and clinically relevant technique for early detection of heart disease, with implications for telemedicine, portable diagnostic systems, and smart stethoscopes. Its organization ensures that it can assist clinicians in early detection of cardiac anomalies without maintaining high computational and operational overhead.

3. Limited GeneralizationAcross Populations

The model may perform less effectively on populations or heart conditions that were underrepresented in the training dataset.

4. False Positives and Negatives

While PulseCheck can detect abnormal heart sounds, it may occasionally classify normal sounds as abnormal (false positives) or miss subtle abnormalities (false negatives).

5. Not a Replacement for Clinical Diagnosis

PulseCheck is a supportive tool for screening but cannot replace comprehensive medical evaluation or expert cardiologist judgment.

VII. CONCLUSION

This research introduces PulseCheck, a deployable and lightweight machine learning system for the automatic classification of PCG signals as Normal orAbnormal. Through robust preprocessing, all-around feature extraction, class balancing, and XGBoost classification, the system is highly accurate while being computationally efficient and deployable in real-time.

While PulseCheck demonstrates good performance, there are a couple of areas where the system can still be enhanced, both in terms of accuracy and field deployment. The areas for the future work can be covered under various areas:

1. Hybrid Model Strategies:

Merger of light classifiers with deep learning feature extractors may be able to learn more complex patterns in PCG signals, particularly rare or weak abnormal states.Combining the explainability of XGBoost with hierarchical feature learning of CNNs or Transformers can provide both accuracy and explainability.

2. Real-Time Embedded Deployment:

Optimizing the system for deployment on handheld devices, smart stethoscopes, and mobile health applications can enable real-time cardiac screening in rural or low-resource environments.Hardware-efficient implementation (e.g., edge computing or FPGAacceleration) can reduce latency and power consumption.

3. Larger and More Varied Datasets:

The major contributions of PulseCheck are: With the addition of other databases in different populations, reprPage 11 of 13 - Integrity Submission are normalized and recording machines, and clinical pSubmission ID trn:oid:::2945:321429872

purified to remove ambient artifacts and ensure uniformity in generalization can be improved.More extended recordings of

al signal will furthert increase sensitivity to atypical but clinically significant occurrences.

4. ExplainableAI Integration:

Training interpretableAI techniques will allow clinicians to understand why the model has identified a signal as abnormal, building confidence and adoption in clinical use.Visualizations of the contribution of features, attention maps, or decision rules can be added to the predictive output.

5. Continuous Learning andAdaptation:

Adding online learning mechanisms might enable PulseCheck to refine and update its model based on new arriving data and sustain high-performance overtime.This can also tailor the system to individual patient groups or device types.

6. Multi-Modal Integration:

Blending PCG signals with other physiological information, like ECG, pulse oximetry, or blood pressure, can improve the accuracy of cardiac screening and deliver more meaningful clinical information.

7. Clinical Validation:

**14** L required t demonstrate e

workability a reliability system real-world healthcare environments.Iterative refinement can be steered through clinicians' feedback, so that the system meets realistic clinical requirements.

Short-term, there is great room for the future of PulseCheck in technology upgrade, integration with clinical practice, and data expansion. These will make it even more a cost-effective, accessible, and credible method of detecting early cardiac disease, potentially redefining telemedicine and portable diagnostic technology.

[1] H. Zhang, Y. Li, and C. Wang, nPCGmix:2A5:Data872 Augmentation Method for Heart Sound Classification", IEEE Journal of Biomedicaland Health Informatics, vol. 28, np. 3. pp. 1-8.2024,

[2] S. Potes, S. Parvaneh, A. Rahman, and B. Conroy, Ensemble of Feature-Based and Deep Learning-Based Classifiers for Detection of Abnormal Heart Sounds" Journal of Biomedical and Health Informatics, vol, no. 5, pp. 1235 1244, 2017.

[3] S. Maknickas and A. Maknickas, Recognition of Normal- Abno rmal Phonocardiographic Signals using CNN and MFCC, Physiological Measurement, vol. 38, no. 8, pp. 1671 1684, 2017.

[4] A. Kay and J. C. Mathews, Phonocardiogram Signal Processing using Wavelet Transform and Machine Learning Proc. IEEE Int. Conf. Biomedical Engineering, pp. 215-220,2020.

[5] D. Kumar and M. Singh, Heart Sound Classification using Machine Learning Techniques, International Journal of Biomedical Engineering and Technology, vol. 39, no. 4, pp. 320 331,2022.

[6] S. Roy, P. Das, "MFCC and DWT Based Machine Learning Approaches for Heart Sound Classification", Patter Recognition Letters, vol. 158, pp. 14-21, 2022.

[7] V. Reddy, T. Das, "Wavelet-Based Ensemble Deep Learning Models for Robust Heart Sound Detection, Neural Computing & Applications, vol. 34, по.9, pp. 20835-20845, 2022.

[8] K. Ahmed, R. Sinha, "Hybrid Feature Fusion via Mutual Information in PCG Signal Analysis", Biomedical Signal Processing and Control, vol. 80, 104216, 2023.

[9] N. Verma, D. Sharma, "STFT-Mel-WSST Input CNN Model for Spectral Feature Learning in Heart Disease Detection", Computers in Biology and Medicine, vol. 140, 105462, 2022.

VIII FUTURE ENCHANCEMENT [10] Y. Takahashi, J. Sun, 'Conv-Transformer Encoder for Local and Global Pattern Detection in Heart Sound Data",

Future directions for advancing PulseCheck highlight several Biosignal Engineering Journal, vol. 27, no. 3, pp. 321-promising areas. Hybrid models that combine deep learning 329,2025

feature extractors with lightweight classifiers could enhance

detection iperformance, particularly for rare or subtle cardiac Classifier Approaches for Heart Sound Screening, Artificial where optimizing the pipeline for embedded devices, smart

sstyesttheomscmopoerse, aocrcmesosibbillee ahnedaltphraacptipclaiclaitnioncslinwicoaulldanmdarkeemothtee [H1P2S]SS. aKndauDr,eMep. AGNillN, "fHoeraHrtaSrmouonndicCFlaesastuifriecaEtioxntraUctsiionng,

settings. Additionally, integrating explainable AI methods can Sensors and Actuators B: Chemical, vol. 343, provide clinicians with clear, interpretable explanations for 130100,2021.

classifications, thereby fostering trust and facilitating adoption

in healthcare practice. Expanding the training datasets to [13] M. Han, Z. Chen, "CWT vs Chirplet: Comparative ipnocpluudlaetionmsoreanddivercsoendPitCioGns rewcoorudlidngs fuarctrhoesrs dimiffperroevnet CTilmases-iFfirceaqtiuoenn"c,y BiomTeradnicsafolrmEnsginefeorirng LHeettaerrts, voSlo. u1n0d,

generalization and robustness. Altogether, PulseCheck

emerges as a deployable, realistic, and accurate solution for [14] X. Xu, Y. Zhao, "Explainable CNN Approaches for **2** PCG Signal Classification", IEEE Journal of Biomedical

t potential significantly impact cardiovascular healthcare and Health Informatics, vol. 27, no. 1, pp. 76-84, 2023

through timely intervention and better patient outcomes. [15] R. Das, S. Singh, "ANFIS Combined with ABC Page 12 of 13 - Integrity Submission Optimization for Lightweight, Submission ID trn:oid:::2945:321429872

IX. REFERENCES

on, Soft1Computing, vol. s27, no. 6, pp. 12456- Submission ID trn:oid:::2945:321429872 12463, 2022.

Page 13 of 13 - Integrity Submission 70 Submission ID trn:oid:::2945:321429872

CHAPTER 10

REFERENCES

[1] H. Zhang, Y. Li, and C. Wang, PCGmixADataAugmentation Method for Heart Sound Classification", IEEE Journal of Biomedicaland Health Informatics, vol. 28, np. 3. pp. 1-8.2024,

[2] S. Potes, S. Parvaneh, A. Rahman, and B. Conroy, Ensemble of Feature-Based and Deep Learning-Based Classifiers for Detection ofAbnormal Heart Sounds" Journal of Biomedical and Health Informatics, vol, no. 5, pp. 1235 1244, 2017.

[3] S. Maknickas andA. Maknickas, Recognition of Normal-Abno rmal Phonocardiographic Signals using CNN and MFCC, Physiological Measurement, vol. 38, no. 8, pp. 1671 1684, 2017.

[4]A. Kay and J. C. Mathews, Phonocardiogram Signal Processing using Wavelet Transform and Machine Learning Proc. IEEE Int. Conf. Biomedical Engineering, pp. 215-220,2020.

[5] D. Kumar and M. Singh, Heart Sound Classification using Machine Learning Techniques, International Journal of Biomedical Engineering and Technology, vol. 39, no. 4, pp. 320 331,2022.

[6] S. Roy, P. Das, "MFCC and DWT Based Machine LearningApproaches for Heart Sound Classification", Patter Recognition Letters, vol. 158, pp. 14-21, 2022.

[7] V. Reddy, T. Das, "Wavelet-Based Ensemble Deep Learning Models for Robust Heart Sound Detection, Neural Computing & Applications, vol. 34, по.9, pp. 20835-20845, 2022.

[8] K.Ahmed, R. Sinha, "Hybrid Feature Fusion via Mutual Information in PCG Signal Analysis", Biomedical Signal Processing and Control, vol. 80, 104216, 2023.

[9] N. Verma, D. Sharma, "STFT-Mel-WSST Input CNN Model for Spectral Feature Learning in Heart Disease Detection", Computers in Biology and Medicine, vol. 140, 105462, 2022.

[10] Y. Takahashi, J. Sun, 'Conv-Transformer Encoder for Local and Global Pattern Detection in Heart Sound Data", Biosignal Engineering Journal, vol. 27, no. 3, pp. 321-329,2025

71

[11] H. Green, D. White, "MFCC Ensemble vs. Single ClassifierApproaches for Heart Sound Screening,Artificial Intelligence in Medicine, vol. 110, 101963, 2021.

[12] S. Kaur, M. Gill, "Heart Sound Classification Using HPSS and DeepANN for Harmonic Feature Extraction, Sensors andActuators B: Chemical, vol. 343, 130100,2021.

[13] M. Han, Z. Chen, "CWT vs Chirplet: Comparative Time-Frequency Transforms for Heart Sound Classification", Biomedical Engineering Letters, vol. 10, pp. 455-462, 2020.

[14] X. Xu, Y. Zhao, "Explainable CNNApproaches for PCG Signal Classification", IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 76-84, 2023

[15] R. Das, S. Singh, "ANFIS Combined withABC Optimization for Lightweight, Flexible Heart Sound Classification, Soft Computing, vol. 27, no. 6, pp. 12456-12463, 2022.

72