APRIL 2025

ST 3082

FINAL PROJECT ANALYSIS

Prepared by :

Group 08

Kavindu Jayawardana s16035
Dinithi Gunarathna s16029
Keshana Nishshanka s16163

Abstract

Cardiovascular diseases continue to be a leading cause of mortality globally, emphasizing the urgent need for timely arrhythmia detection. This study introduces a robust arrhythmia classification system developed using high-quality ECG data sourced from the MIT-BIH Arrhythmia Database. The system features a comprehensive machine learning pipeline that includes advanced preprocessing, targeted feature extraction, and the application of sophisticated algorithms such as random forest, XGBoost, and decision trees. Specifically designed for Sri Lanka's public healthcare sector, this methodology addresses the limitations of consumer-grade IoT devices by providing rapid, accurate, and non-invasive heartbeat classification. The system not only enhances early arrhythmia detection but also presents a cost-effective solution in the face of financial constraints, resource shortages, and high patient volumes, ultimately supporting improved patient outcomes in challenging clinical settings.

Table of Contents

List of Tables

List of Figures

Introduction

Cardiovascular diseases (CVDs) are the leading global cause of death, as reported by the World Health Organization (WHO). They often arise from undetected cardiac arrhythmia, which can lead to heart failure if not addressed. Arrhythmia disrupts the heart's ability to pump blood effectively, causing fluctuations in heart rate that patients may not always recognize. Therefore, automatic detection of arrhythmia is vital to prevent severe complications associated with CVDs. Continuous monitoring of ECG signals is essential for identifying different types of CVDs. (Sakib et al., 2021)

The traditional method for long-duration ECG monitoring is recognized as invasive, time-consuming, and costly, often restricting the subject's daily activities during the ECG data collection process. In developing countries like Sri Lanka, where financial constraints, limited medical resources, and long queues in government hospitals are common, these challenges become even more acute.

In response to these challenges, our study aims to develop an efficient multi-class heartbeat (HB) classification system designed specifically to support medical practitioners in clinical settings in Sri Lanka. By leveraging advanced machine learning techniques, our model can rapidly and accurately identify various heartbeat patterns, enabling timely and informed clinical decision-making—a critical need when long-term cardiac monitoring is not always feasible. In a healthcare system burdened by high demands on medical professionals and long patient queues, an accurate and time-efficient tool for predicting arrhythmia can aid in the early evaluation of heart conditions, ultimately benefiting both clinicians and patients.

To enhance practical clinical use, our proposed system integrates with a medical-grade ECG monitor, specifically the Mindray BeneView T5, which captures lead II and lead V raw signals in .dat format for further analysis. Rather than simply classifying patients as normal or abnormal, our objective is to stratify patients into five superclasses that reflect the severity of their condition. By employing a robust, medically certified ECG monitoring system alongside secure, non-cloud-based analysis, our model advances current methods and meets the unique needs of a resource-constrained healthcare environment, ultimately contributing to improved outcomes in the early detection and management of arrhythmia

Description of the Dataset

To develop our arrhythmia classification model, we utilized the MIT-BIH Arrhythmia Database from PhysioNet, a crucial resource for ECG research since 1980. This database comprises 48 half-hour ECG recordings, sampled at 360 Hz from 47 subjects, using both lead II and lead V. The recordings are divided into two subsets: the 100 series, which includes 23 subjects from over 4,000 Holter recordings, and the 200 series, which features 25 subjects with less common arrhythmias.

Each ECG record is labelled with one of five heartbeat classes according to the AAMI EC57 standard: Normal (N), Supraventricular ectopic (S), Ventricular ectopic (V), Fusion (F), and Unknown (Q). The Normal class is the most prevalent, with more than 75,000 samples.

We employed a pre-processed version of the dataset that features 34 extracted characteristics for each heartbeat—17 derived from lead II and 17 from lead V5. These features encompass RR intervals, duration intervals, amplitude peaks, and morphology traits. Each entry includes a “record” identifier for the patient and a “type” column for heartbeat classification.

Methodology

This study proposes a machine learning-based solution for the early detection and classification of arrhythmias using ECG signals obtained from clinical-grade monitoring devices.

Initially, the MIT-BIH Arrhythmia Database was examined for data quality issues, confirming no duplicates or missing values were present, with minimal outliers observed. The dataset was split into training (80%) and testing (20%) sets using stratified sampling based on the arrhythmia type to maintain class distribution. Due to class imbalance, the dominant normal (N) class was downsampled to 30,000 samples in the training set, and Synthetic Minority Over-sampling Technique (SMOTE) was applied to increase the representation of minority classes.

Feature engineering included applying centering and scaling to all features. Multiple classification algorithms were evaluated, including Random Forest, XGBoost, Decision Trees, Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM). Performance was assessed through accuracy, macro F1 score, and class-specific metrics to determine the optimal model for arrhythmia classification.

Arrhythmia Classification Model

Data Acquisition

The raw ECG data is collected using the Mindray BeneView T5 medical-grade multi-parameter monitor. This device supports multiple ECG leads, including Lead II and Lead V, which are critical for analyzing cardiac arrhythmias. It stores ECG recordings in .dat format, which is suitable for long-term monitoring. The use of this device ensures high reliability and accuracy, characteristics often lacking in IoT-based consumer solutions currently under research.

Brand	Model Series	Type	Lead Support	Common Use
Mindray	BeneView T5/T8	Multi-parameter	3, 5, 12	ICU, ER

Figure 1: Multi-parameter monitor

Preprocessing of Raw Signals

The raw ECG signals acquired in .dat format undergo preprocessing using R programming tools. This process includes noise reduction, signal normalization, and segmentation to prepare the data for machine learning analysis. The preprocessing step has been tested and optimized with the assistance of artificial intelligence techniques. The signals are converted from .dat to .csv format, allowing for structured analysis and feature extraction. Observations are collected over 1 to 2 minutes, during which a patient can produce between 50 to 250 heartbeat samples, each treated as a single instance for classification.

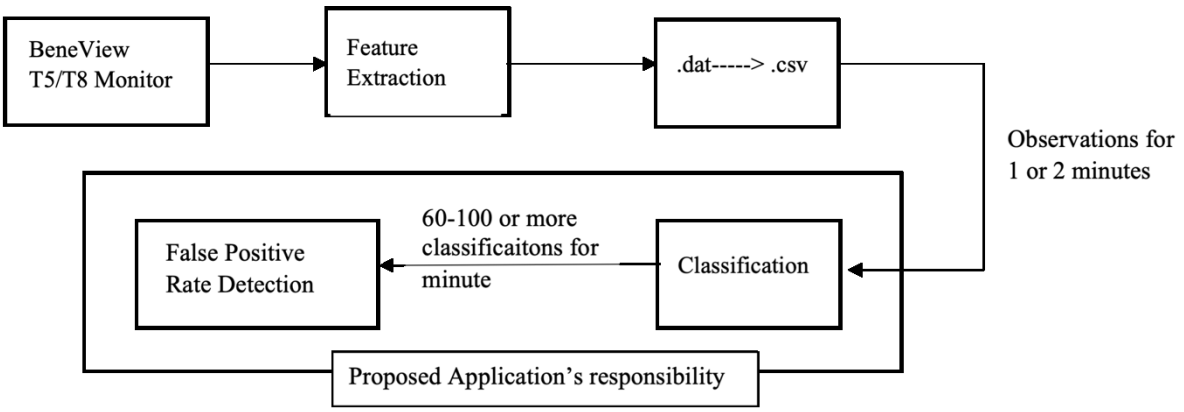


Figure 2: ECG classification model workflow

Feature Extraction

Following preprocessing, essential features are extracted from the ECG signals. These features are derived from the Lead II and Lead V signals and include:

- RR Interval Features: Average RR, RR, Post RR

- Heartbeat Interval Features: PQ Interval, QT Interval, ST Interval, QRS Duration

- Amplitude Features: P peak, Q peak, R peak, S peak, T peak

- Morphology Features: QRS Morphology Features (0 to 4)

These 17 features per lead (34 in total) provide comprehensive information about the heart's electrical activity, critical for identifying arrhythmic patterns.

Classification Pipeline

As shown in the process pipeline, the extracted features are input into a machine-learning model that performs multi-class classification. The model predicts the heartbeat type using five superclass labels from the AAMI EC57 standard

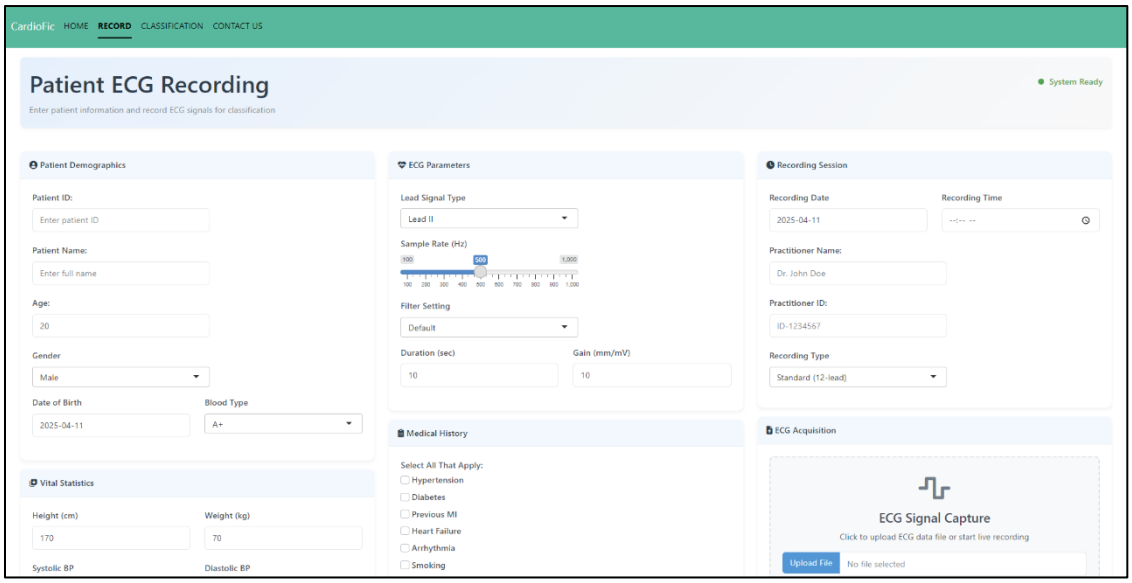
Each superclass corresponds to a set of annotations defined by the MIT-BIH Arrhythmia Database. The classification model processes 60 to over 100 heartbeats per minute, allowing for near real-time predictions. It also includes a False Positive Rate Detection component to ensure reliability in a clinical environment.

Table 1: Mapping the MIT-BH Arrhythmia dataset heartbeats to the AAMI heartbeat classes.

Heartbeat Super Class	Heartbeat Annotation
N (Normal)	N (Normal)
	L (Left bundle branch block beat)
	R (Right bundle branch block beat)
	e (Atrial escape beat)
	j (Nodal (junctional) escape beat)
	A (Atrial premature beat)
	a (Aberrated atrial premature beat)
S (Supraventricular ectopic beat)	J (Nodal (junctional) premature beat)
	S (Supraventricular premature beat)
	V (Premature ventricular contraction)
V (Ventricular ectopic beat)	E (Ventricular escape beat)
	F (Fusion of ventricular and normal beats)
F (Fusion beat)	Q (Unclassifiable beat)
	/ (Paced beat)
Q (Unknown beat)	f (Fusion of paced and normal beat)

Application Deployment

An R Shiny-based application was developed to deploy the ECG classification model. It allows users to upload .csv files with 32 named features, processes up to 200 heartbeats per session, and identifies the dominant class based on majority predictions. The app provides visual and statistical reports, including class distribution, confidence scores, and individual results—making it a practical, real-time tool for clinical and research use.



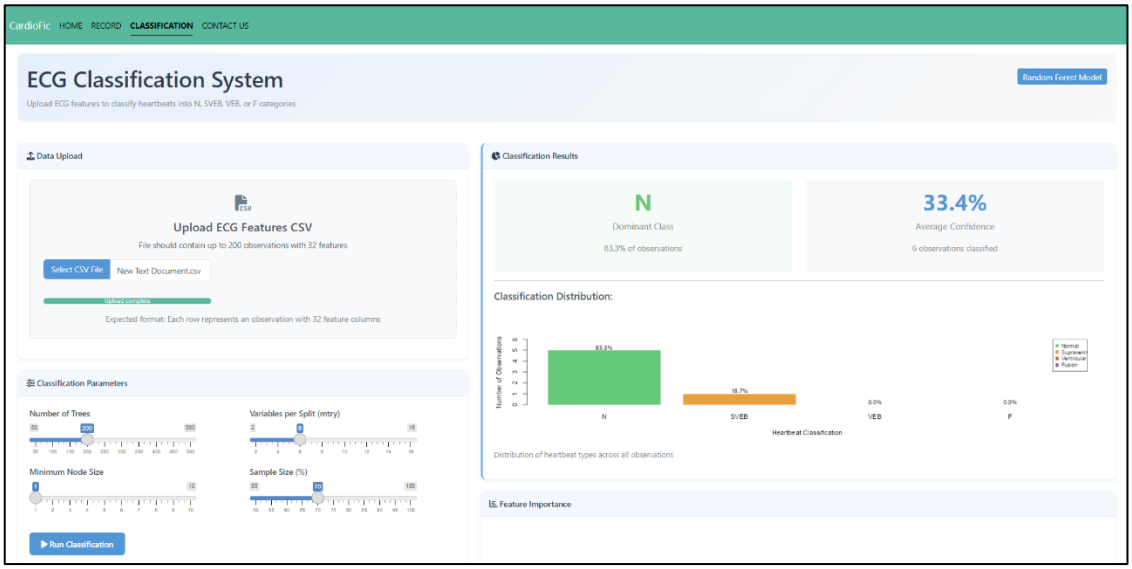


Figure 3(a) and Figure 3(b): UI of the classification Model

Results

Exploratory Data Analysis

The exploratory analysis showed that Ventricular Ectopic Beats (VEB) have a significant and irregular impact on ECG morphology compared to other heartbeat types. VEBs are often linked to severe cardiovascular issues, resulting in marked deviations from normal sinus rhythms, characterized by unpredictable morphology and abnormal RR intervals.

In contrast, Supraventricular Ectopic Beats (SVEB) exhibit more stable morphology, with key distinguishing features in RR interval patterns. The analysis indicates that SVEBs are closely related to changes in the ratio of previous to post RR intervals, supporting the clinical view that they originate from above the ventricles and influence rhythm without severely distorting the waveform.

Advanced Analysis

Comparison of Models

Table 2: Performance Comparison of the machine learning models before and after employing SMOTE for Test Data

	Without Applying SMOTE		After Applying SMOTE	
Model	Accuracy	Precision	Accuracy	Precision
RF	98.58%	90.820%	98.86%	94.524%
XGB	99.04%	96.08%	95.77%	76.53%
LDA	90.98%	63.96%	79.15%	35.01%
SVM	96.46%	74.71%	98.6%	91.721%
Decision Tree	97.27%	88.49%	91.81%	62.86%

Five classification models were evaluated with and without SMOTE. The Random Forest (RF) model achieved 98.86% accuracy with SMOTE and 98.58% without it, showing stable performance. The XGBoost (XGB) model had a higher accuracy of 99.04% without SMOTE, but dropped to 95.77% with it, also showing a precision of 76.53%. The Decision Tree and LDA models performed poorly, especially after applying SMOTE. In contrast, the SVM model maintained strong accuracy at 98.6% with SMOTE.

Feature Importance

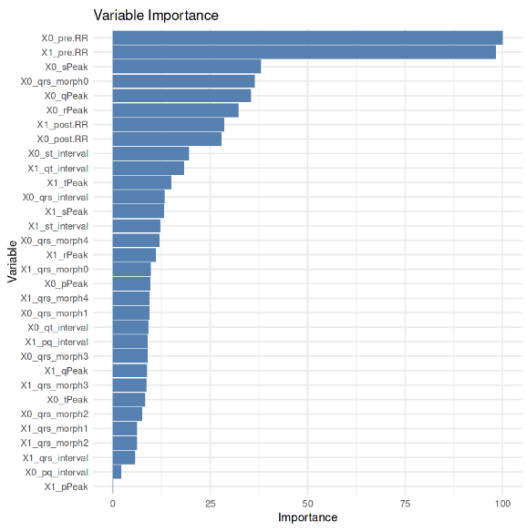


Figure 4: Variable Importance Graph of RF Model (scaled)

The variable importance plot shows that the interval between the current and previous R peak is the key predictor for distinguishing arrhythmia classes. Features related to s-peak amplitude and certain morphological characteristics are also significant. However, QRS morphology features rank lower in importance.

SMOTE

SMOTE was achieved by improving the sampling fractions of “VEB,” “SVEB,” and “F” by 2, 3, and 3 times, respectively.

Table 3: Number of observations in each class before and after employing SMOTE

Before SMOTE					After Smote				
F	N	Q	SVEB	VEB	F	N	Q	SVEB	VEB
643	30000	12	2224	5608	1929	30000	12	6672	11216

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction     F     N     Q  SVEB   VEB
      F      127     4     0     0     2
      N       17 17897     3    50    16
      Q        0     0     0     0     0
      SVEB     0    53     0   501     2
      VEB     16    62     0     4  1381

Overall Statistics

               Accuracy : 0.9886
                 95% CI : (0.9871, 0.99)
    No Information Rate : 0.8948
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9417
```

Figure 5: Confusion Matrix

The analysis of the confusion matrix confirms that our model effectively distinguishes between the three main classes: Normal, Ventricular Ectopic Beats (VEB), and Supraventricular Ectopic Beats (SVEB), which are essential for clinical decision-making.

Discussion

The results indicate that our arrhythmia classification system, trained with a blend of original and SMOTE-adjusted datasets, achieved exceptional performance across various machine learning models. Among those evaluated, the Random Forest (RF) classifier consistently surpassed others in terms of precision and accuracy, attaining a training accuracy of 100% and a test accuracy of 98.86% with SMOTE. Although SMOTE provided only a modest improvement of around 0.028%, it effectively mitigated class imbalance, particularly for the minority arrhythmia classes. While XGBoost (XGB) achieved a slightly higher accuracy of 99.04% without SMOTE, its performance significantly declined after applying SMOTE (with a 95.77% accuracy and 76.53% precision), indicating a tendency toward overfitting. The Decision Tree and LDA models displayed lower overall accuracies, especially after balancing. The SVM model maintained robust performance post-SMOTE with an accuracy of 98.6%, but it did not demonstrate the same level of robustness as RF. Given its low tendency towards overfitting, consistent precision the Random Forest model was ultimately chosen as the final classifier for the proposed application.

The importance plots highlighted that the interval between current and previous R peaks is crucial for distinguishing arrhythmia classes. Although QRS complex deformation features are influential, further analysis reveals that their information is largely redundant, as it is captured by amplitude and interval measurements. Thus, removing most QRS morphology features could simplify the data and reduce computational costs without significantly impacting classification performance.

The confusion matrix demonstrates that the model effectively differentiates between Normal, VEB, and SVEB classes, which is especially important due to the risks associated with misclassifying VEB. While 16 observations were incorrectly labeled as Normal, the model's impressive accuracy of 98.86% indicates its potential for clinical use, particularly when supported by expert evaluation. Overall, it can serve as a reliable decision support tool in conjunction with clinical judgment.

The proposed system is designed for real-world clinical settings, where patients can be monitored up to two minutes, yielding 100 to 200 heartbeat observations per session. Final classification of overall pulses is made through a majority vote, ensuring reliable predictions avoiding the risk of misclassification if single observation is used. The probability of misclassification of 50% or more out of 10 classifications is around 6×10^{-8} which shows **near 100% accuracy per single patient.**

Given the limited occurrence of the Unknown (Q) category and the modest improvements brought by using SMOTE, our model primarily focuses on classifying Normal, VEB, and SVEB beats, reflecting clinical priorities in Sri Lankan government hospitals, where resource constraints and high patient volumes necessitate rapid and cost-effective diagnostic tools. Ultimately, the random forest model offers a practical solution for continuous, real-time arrhythmia detection in clinical practice.

Conclusion

This study presents a robust arrhythmia classification system specifically designed for public healthcare sector in Sri Lanka. By utilizing high-quality ECG data and advanced machine learning techniques, our approach addresses the limitations of existing research that often relies on consumer-level IoT devices, which may be neither reliable nor accessible in Sri Lanka. Above strategy not only enhances early arrhythmia detection but also optimizes resource utilization under financial and operational constraints, ultimately fostering better clinical decision-making and improving patient outcomes in settings with limited access to long-term cardiac monitoring. In contexts where financial limitations, resource scarcity, and high patient volumes restrict long-term cardiac monitoring, our system offers swift, precise, and non-invasive heartbeat classification, thereby significantly improving early arrhythmia detection and patient outcomes.

Technical Appendix and R Code

Link for the dataset: MIT-BIH Arrhythmia Database

The R code used in our project is conveniently accessible through our GitHub repository via this link: https://github.com/keshanagn/St-3082-DA-Final-Project

References

Sakib, S., Fouda, M. M., & Fadlullah, Z. M. (2021). Harnessing Artificial Intelligence for Secure ECG Analytics at the Edge for Cardiac Arrhythmia Classification. In CRC Press eBooks (pp. 137–153). https://doi.org/10.1201/9781003028635-11

World Health Organization: WHO. (2021, June 11). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Wagner, G. S. (2008). Marriott's practical electrocardiography (11th ed.). Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). Model-based clustering and classification for data science. Cambridge University Press.

Sakib, S., Fouda, M. M., Fadlullah, Z. M., Nasser, N., & Alasmary, W. (2021). A Proof-of-Concept of Ultra-Edge smart IoT sensor: a continuous and lightweight arrhythmia monitoring approach. IEEE Access, 9, 26093–26106. https://doi.org/10.1109/access.2021.3056509