CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

APPLICATION OF MACHINE LEARNING
FOR SINGLE-LEAD ECG-BASED ARRHYTHMIA CLASSIFICATION
VIA SMART WEARABLE DEVICES

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Software Engineering

By
Himanshu Kumar

December 2023

The thesis of Himanshu Kumar is approved:

Xunfei Jiang, Ph.D. Date

Robert D. Mclhenny, Ph.D. Date

Taehyung Wang, Ph.D., Chair Date

California State University, Northridge

Acknowledgements

The journey to complete this thesis has been challenging yet enriching, both academically and personally. I am deeply grateful for the invaluable support and guidance of those who have been instrumental in reaching this milestone.

Foremost, I express my deepest appreciation to my thesis advisor, Dr. Taehyung Wang. His insightful guidance and constructive feedback have been vital throughout my research journey. I am truly thankful for his patience, expertise, and encouragement.

I am also deeply grateful to Dr. Xunfei Jiang, the lead faculty member at the helm of the Smart Wearables Project within ARCS. Her expertise and steadfast support have been crucial to the project's success and, by extension, my achievements within it.

Special thanks go to ARCS, the Autonomy Research Center for STEAHM funded by NASA, which provided the opportunity to work on the Smart Wearables Project. Working alongside fellow students from diverse fields of study has enriched my understanding and approach to this project, and I am grateful for their collective brilliance and contribution.

Finally, I would like to thank my family, who deserves particular recognition for their unwavering love and support. Their constant belief in me and my aspirations has been the bedrock of my resilience, especially during the most demanding stages of my research.

Table of Contents

List of Figures

List of Tables

Abstract

Application of Machine Learning
for ECG-based Arrhythmia Classification
via Smart Wearable Devices

By

Himanshu Kumar

Master of Science in Software Engineering

Cardiovascular diseases, notably arrhythmias, are a leading cause of mortality worldwide. The integration of smart wearable devices into healthcare has opened new avenues for electrocardiogram (ECG) data monitoring and analysis. Despite advancements, most studies have either focused on detecting Sinus Rhythm (SR) and Atrial Fibrillation (AF) using single-lead ECG from wearables or have relied on conventional 12-lead ECG data.

This research transcends these limitations by exploring the capabilities of single-lead ECG sensors for a more comprehensive arrhythmia classification. Utilizing the MIT-BIH Arrhythmia Database from Physionet.org, this study introduces a novel approach by adapting the original two-lead ECG data to emulate the single-lead data acquisition of modern smart wearables.
A Random Forest (RF) machine learning model was developed and trained on this adapted dataset for the classification of normal beats, supraventricular ectopic beats, and ventricular ectopic beats.

The third experiment yielded significant results, with the RF model achieving an overall accuracy of 95.69% and a weighted average F1 score of 95.37%. Specifically, the model demonstrated a precision of 95.94%, recall of 97.73%, and F1-score of 96.82% for normal beats (N). For supraventricular ectopic beats (S), the precision was 45.22%, recall 37.62%, and F1-score 41.07%. Ventricular ectopic beats (V) were classified with a precision of 92.6%, recall of 86.68%, and F1-score of 89.62%. The sensitivity, specificity, positive predictivity, false positive rate, and accuracy for each class further underscore the model's efficacy.

These results underscore the potential of RF models in classifying a broader range of arrhythmias using single-lead ECG data. This research not only contributes significantly to the understanding and application of single-lead ECG in smart wearable technology but also aims to provide valuable health insights for athletes at California State University as part of the Smart Wearables Project at the NASA-funded Autonomy Research Center for STEAHM (ARCS).

Chapter 1 - Introduction

In today's world, where technology and healthcare increasingly intersect, smart wearables stand out as a critical innovation, particularly in athletic performance and medical diagnostics. The global market for wearable healthcare technology, encompassing devices such as smartwatches and activity trackers, is on a rapid growth trajectory, projected to reach nearly US$200 billion by 2027 [1]. Technological advancements in smart wearables have become a staple in everyday life. Consumer-grade devices such as smartwatches, rings, wristbands, and shirts are increasingly utilized for health management, thanks to their advanced sensors and high processing power [4].

In both the athletic industry and the broader world, the demand for real-time biometric data analysis is increasing. The integration of ECG sensors into wearable devices has become crucial for monitoring cardiac health. These smart wearables facilitate the acquisition of ECG data in various settings, aiding in prevention and risk management. Consequently, the early detection and classification of arrhythmias, which are irregular heart rhythms that are often undetected and can lead to sudden cardiac incidents, have become essential. This is particularly true for athletes, who may face a higher risk of cardiac issues due to the intense physical exertion and stress their bodies endure [4].

However, the in-depth analysis of ECG data for detecting common types of arrhythmias, especially using single-lead ECG sensors typically found in consumer wearables, remains relatively unexplored. Although these sensors are well-suited for detailed ECG data collection, their full potential in arrhythmia detection has not been fully realized. This limitation is partly due to the fact that single-lead ECGs provide only one view of the heart’s electrical activity, compared to the comprehensive 12 different electrical views offered by standard 12-lead ECGs. While Random Forest models have been employed to detect conditions like Sinus Rhythm (SR) and Paroxysmal Atrial Fibrillation (PAC) [6] using single-lead ECGs, there is still a significant gap in utilizing this approach for a broader spectrum of arrhythmias.

This thesis is integral to the Smart Wearables Project [3] at the Autonomy Research Center for STEAHM (ARCS) [2] at California State University, Northridge (CSUN). This initiative represents a collaborative effort across multiple departments, leveraging smart textile technology with built-in electronics and microchips to analyze crucial health data from athletes, with the goal of enhancing their well-being through data-driven insights.

Specifically, this thesis aims to leverage the Random Forest machine learning model to classify normal beats, supraventricular ectopic beats, and ventricular ectopic beats using data from single-lead ECG sensors. The goal is to distinguish between the normal electrical activity originating from the sinus node and the ectopic activity originating elsewhere. While ectopic beats are often benign and not indicative of immediate health issues, their detection and frequency quantification are significant in prevention and risk management. Furthermore, the early detection of ventricular beats is particularly important, as it can serve as a predictor of heart failure and mortality risk [8].

This study aims to propose a Random Forest model designed to interpret data from single-lead ECG sensors, as featured in smart wearables like the Apple Watch, Samsung Watch, and the Hexoskin ProShirt. The primary objective is to extend the model's application beyond the detection of Sinus Rhythm and Paroxysmal Atrial Fibrillation by focusing on the classification of supraventricular ectopic beats and ventricular ectopic beats. This approach acknowledges the practical constraints and opportunities within the evolving landscape of wearable technology and aims to enhance the utility of single-lead ECG data in improving health outcomes for athletes and individuals with a higher risk of cardiac conditions.

Chapter 2 - Related Works

The advent of smart wearable devices has opened new possibilities in cardiac health monitoring. Among these, single-lead ECG sensors embedded in consumer wearables, such as the Hexoskin ProShirt and Apple Watch, have become a focal point for research. These devices have been effectively used to detect conditions such as Atrial Fibrillation (AF) and Sinus Rhythm (SR) [7]. However, their capacity to identify a wider range of cardiac anomalies is still an area ripe for exploration. This Literature Review examines existing studies on classification of arrhythmia and related cardiac conditions through ECG data, highlighting the innovative nature of this thesis.

The application of smart wearable devices to classify a broader spectrum of arrhythmias, beyond the binary categorization of normal or abnormal rhythms, is not extensively documented. This gap is primarily due to the inherent limitations of single-lead ECG (1L-ECG) data, which offers a limited perspective compared to the comprehensive view provided by standard 12-lead ECGs used in hospitals.

There have been notable studies utilizing machine learning for arrhythmia classification, predominantly with data from 12-lead ECGs. For instance, a study applied a 12-layer 1-d Convolutional Neural Network (CNN) to categorize 12-lead ECG signals into five distinct categories, including SR, LBBB, RBBB, PVC, and Atrial Premature Beat [10]. Their method achieved impressive results, with high positive predictive value, sensitivity, and F1 score, all exceeding 0.976. This success demonstrates the effectiveness of machine learning in heart disease identification using 12-lead ECG data.

In contrast, another study utilizing single-lead ECG data introduced an algorithm to differentiate between ischaemic and non-ischaemic ST changes employing a decision tree model [9]. This model attained an accuracy of 90% for ischaemic episodes and 86% for non-ischaemic episodes, illustrating the potential of machine learning in refining 1L-ECG analysis.

Furthermore, a distinct study implemented a Random Forest model to classify various arrhythmias and heart-related conditions, using morphological features extracted from a 2-lead ECG signal. This approach achieved an overall accuracy of 96% [11], highlighting the Random Forest model's capability in arrhythmia classification.

The distinctive contribution of this thesis lies in its effort to enhance the Random Forest (RF) model for arrhythmia detection using single-lead ECG data, thereby simulating the functionality of consumer-grade smart wearables like the Apple Watch and Hexoskin ProShirt. It focuses on a comprehensive classification approach, aiming to accurately identify and distinguish between normal beats (N), supraventricular ectopic beats (SVEB), and ventricular ectopic beats (VEB).

This research not only seeks to fill a critical gap in medical research but also aims to transform the practical application of smart wearables in athletic health monitoring. By extending the capabilities of single-lead ECGs, traditionally limited in their diagnostic scope, this study broadens their potential applications. The refined capacity of RF models to classify a wider range of arrhythmias from single-lead ECG data could significantly enhance early detection and management of various heart conditions in athletes and the wider population. This approach represents a significant advancement in the integration of machine learning and single-lead ECG data within the field of wearable technology, underscoring its potential to substantially improve cardiac health monitoring.

Chapter 3 - Background Study

This section is devoted to a systematic exploration of the underlying components that form the bedrock of this research. It is divided into subsections that scrutinize the current state of knowledge and technology in each relevant area, providing the reader with a holistic view of the existing landscape and identifying where this research will contribute new insights.

3.1. Research on ECG Data

The objective of this thesis is to contribute to the Smart Wearables Project by employing machine learning techniques to analyze Electrocardiogram (ECG) data, which measures the heart's electrical activity across repeated cardiac cycles [12]. Electrical impulses produced by the heart as it beats are captured by placing electrodes on the skin. These recorded tracings are then displayed as a series of waves on a graph of voltage versus time.

To measure the electrical activity from different angles (or leads) around the heart, several electrodes are attached to specific locations on the skin [12]. As shown in Fig. 3.1, a standard 12-lead ECG uses 10 electrodes (four on the limbs and six on the chest) to produce twelve different electrical views of the heart, thus providing a comprehensive picture of the heart's electrical activity:

- Limb leads (I, II, III)
    - ○ Lead I is the difference in voltage between the left and the right arm.
    - ○ Lead II is the difference in voltage between the left leg and the right arm.
    - ○ Lead III is the difference in voltage between the left leg and the left arm.

- Augmented limb leads (aVR, aVL, aVF)
    - aVR measures the right arm's voltage relative to a combination of the left arm and the left leg.
    - aVL measures the left arm's voltage relative to a combination of the right arm and the left leg.
    - aVF measures the left leg's voltage relative to a combination of both arms.

- Chest or precordial leads (V1, V2, V3, V4, V5, V6)
    - Unipolar leads that measure the voltage at each chest electrode relative to a combination of all limb electrodes.

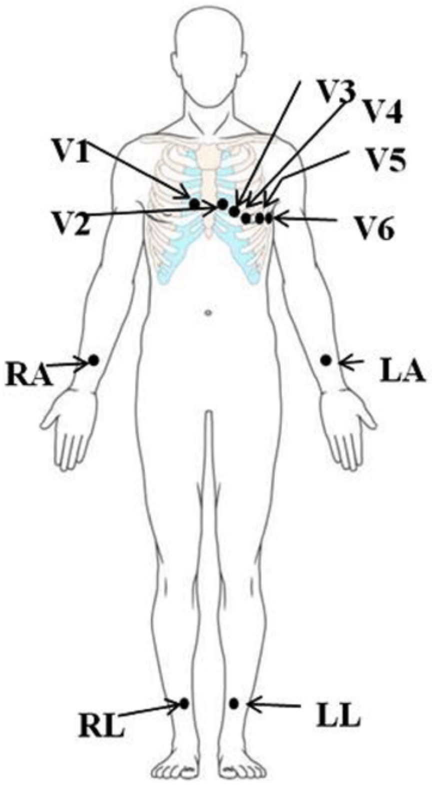


Figure 3.1: Standard 12-lead ECG with 10 electrodes

As shown in Fig. 3.2 below, an ECG tracing consists of waves and segments that represent different phases of a heartbeat [12]:

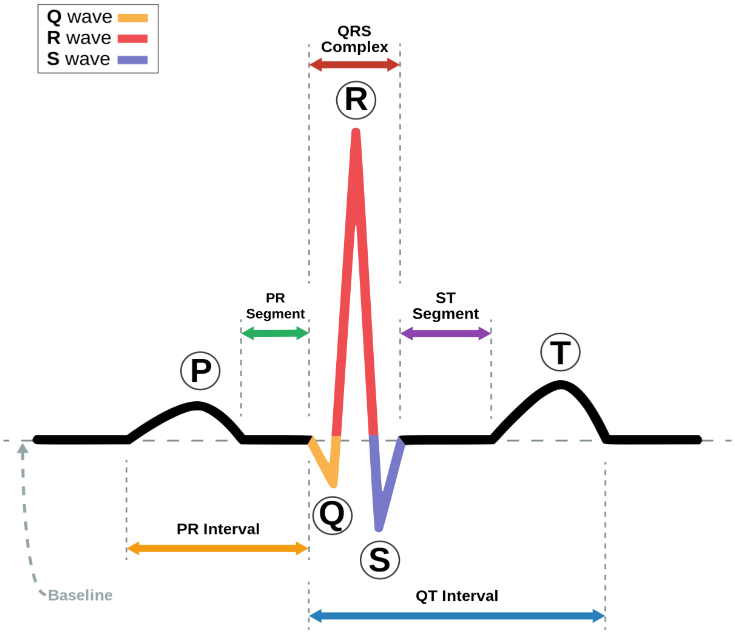


Figure 3.2: ECG tracing with waves and segments

- P Wave:

  Represents atrial depolarization or the electrical activity that causes the atria (upper chambers of the heart) to contract.

- QRS Complex:

  Represents ventricular depolarization or the electrical activity that causes the ventricles (lower chambers of the heart) to contract.

- T Wave:

  Represents ventricular repolarization, which is the period when the ventricles are resetting electrically and preparing for the next contraction.

- U Wave (when present):

  Represents further ventricular repolarization. It's not always seen and is usually small when visible.

- PR Interval:

  The time from the start of the P wave to the start of the QRS complex, representing the time for the electrical impulse to travel from the atria to the ventricles.

- R-R Interval:

  Time between two consecutive R waves in the QRS complex, representing one full cardiac cycle, from one beat to the next.

- QT Interval:

  Time from start of the QRS complex to end of the T wave, representing total time for ventricular depolarization and repolarization.

- ST Segment:

  Represents the period between the end of ventricular depolarization and the beginning of ventricular repolarization.

- Irregularities are identified if intervals are not within these normal ranges:

  - QRS Interval: 80 ~ 120 ms
  - P-R Interval: 120 ~ 200 ms
  - Q-T Interval: 350 ~ 440 ms
  - R-R Interval: 600 ~ 1200 ms
  - T Interval: 160 ~ 280 ms
  - P Interval: 80 ~ 120 ms

The decision to focus on ECG data was made based on the following considerations:

- ECG data is one of the most commonly used types of data in both sports science and health research. This data holds paramount significance in the realm of healthcare and diagnostics, directly relating to an individual's health, cardiovascular fitness, and more.

- While ECG data can be complex, there are established methods for analyzing it, and it's generally easier to work with than some other types of biometric data.

- ECG data lends itself exceptionally well to machine learning (ML) applications. The continuous, high-resolution nature of heart rate data provides rich features for algorithms to analyze.

- ECG data provides a wealth of information about the heart's electrical activity and is vital for a range of diverse applications:

- ECG is instrumental in detecting cardiac anomalies like arrhythmias, coronary artery disease, and other disorders influencing the electrical behavior of the heart.

- Changes in the ECG waveforms can suggest myocardial ischemia (i.e. restricted blood flow to parts of the heart) or myocardial infarction (i.e. heart attack).

- ECG can be used to classify stages of heart failure.

- Some medications can cause changes in the ECG, which can be detected and monitored using ECG data.

- Conditions like atrial or ventricular septal defects may produce changes in the ECG.

3.2. Research on Hexoskin

Following the initiation of the Smart Wearables Project sponsored by ARCS, my role has undergone a subtle transition, leading to a redefinition of my tasks and responsibilities. While the core objective remains the application of machine learning techniques to analyze biometric data, specifically ECG data, the emphasis has shifted.

The primary goal is to provide student-athletes from the CSUN Athletic Department with valuable health insights by analyzing ECG data. One of the smart wearables used to collect this data will be the Hexoskin Pro Kit [6], a specialized smart vest that allows the recording of physiological and movement data, notably including electrocardiogram (ECG) signals.

The Hexoskin Pro Kit, as shown in Fig. 3.3 below, includes the Hexoskin ProShirt and the Hexoskin Smart Recording Device. This smart wearable collects a diverse array of physiological data about the user through its integrated sensors. These include a continuous single-lead electrocardiogram (1L-ECG), accelerometer, gyroscope, respiratory sensor, magnetometer, and skin temperature sensor, collectively capturing a comprehensive range of data points.

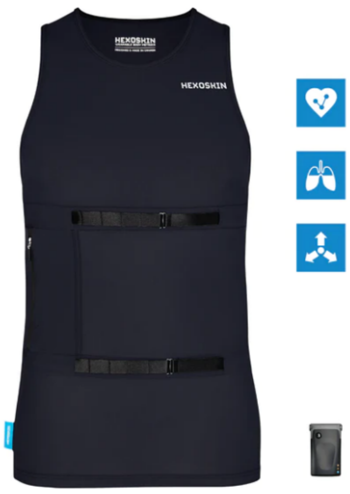


Figure 3.3: Hexoskin Pro Kit

3.2.1. Hexoskin - Data Collected

- ECG Data: heart rate, heart rate variability (HRV) (for stress monitoring, effort, load, and fatigue assessments to prevent overtraining or injuries), QRS events, heart rate zones, heart rate maximum (HR Max), and heart rate recovery (HRR).

- Respiratory Data: breathing rate/frequency, breathing volume, tidal volume, inspiration & expiration events, minute ventilation (L/min), and also VO2 max (maximum volume of oxygen a person can use during intense exercise).

- Accelerometer Data: acceleration, activity intensity (steps, cadence, positions, stride, calories burned, and sleep).

3.2.2. Hexoskin - ECG Data Specifications

- Number of Leads:
  The Hexoskin ProShirt uses a Continuous single-lead ECG (1L-ECG) to collect data of the heart's electrical activity.

- ECG Duration:
  The Hexoskin ProShirt can record ECG data continuously for up to 36 hours. You can also choose to record ECG data in shorter intervals, such as 30 seconds or 60 seconds.

- ECG Sampling Rate:
  The data is sampled at 256 Hz with a 12-bit resolution, meaning that there are 256 data points for each second of ECG data.

- Data Retrieval:
  ECG data can be accessed in segments, with each segment covering any designated time interval. However, there is a restriction of 600 seconds maximum per API request.

- Derived ECG Metrics:
  The Hexoskin API also provides derived ECG metrics, calculated from raw ECG data, such as heart rate, heart rate variability (HRV), ectopic beats, heart rate recovery after exercise, ST segment changes, and atrial fibrillation detection.

- R-peak Field:
  The Hexoskin API also provides an R-peak field to mark the R wave spikes in the ECG waveform, which correspond to heartbeats.

3.2.3. Hexoskin - Analysis of Related Studies

Hexoskin is the industry-leading smart clothing clinically and has been used in over 200 scientific and peer-reviewed publications [6], covering a wide range of fields of study from scientific validations of the Hexoskin cardiac, pulmonary, and activity smart clothing sensors, to various subjects and health applications.

After analyzing these publications, 23 papers were identified related to machine learning applications on biometric data collected using Hexoskin.

Below is a summary of the main findings from these papers:

- Types of Data Analyzed:
  - ECG Data
    - Heart Rate (HR).
    - Heart Rate Reserve (i.e. max HR - resting HR).
    - Δ HR (current HR value - previous HR value).
    - ECG Signals.
    - RR Intervals.

  - Respiratory Data
    - Breathing Frequency/Rate.
    - Minute Ventilation (VE).
    - Tidal Volume (Vt).

- Machine Learning (ML) Models Used on ECG Data:
  - Random Forest (RF).
  - Support Vector Machine (SVM).
  - K-Nearest-Neighbor (KNN).
  - Rotation Forest (RoF).
  - Decision Tree (DT).
  - Temporal Convolutional Network (TCN).
  - Bagged Tree Ensemble (BT).
  - Support Vector Regression (SVR).
  - Multiple Linear Regression (MLR).

- Machine Learning (ML) Applications:
  - Oxygen Uptake (VO2 Max) Prediction
    - ML models used: TCN, RF, SVR
  - Fatigue Detection
    - ML models used: RNN, LSTM, GRU, MLR

- Activity Classification
  - ML models used: TCN, DT, NB, KNN, SVM, NN
- Energy Expenditure
  - ML models used: RF
- Detection of Sinus Rhythm (SR) and Paroxysmal Atrial Fibrillation (PAF) segments in ECG data
  - ML models used: RF, SVM, KNN, RoF, EL
- Extract Kinetics
  - ML models used: RF

3.3. Research on Random Forest

After conducting a comprehensive analysis of Hexoskin-related studies and considering that the objective of this thesis is to apply machine learning on ECG data to uncover valuable health insights, the decision was made to work with a Random Forest machine learning model for the following reasons:

- Overview:
  The Random Forest algorithm is a widely recognized supervised ensemble learning method in machine learning, versatile for both classification and regression problems. This algorithm operates by creating a multitude of decision trees during its training process [13]. In classification tasks, the Random Forest's output is the mode class predicted by the majority of trees. For regression tasks, it computes and returns the mean prediction from all the individual trees.

- Established Efficacy in Hexoskin Research:
  Notably, the Random Forest model is a well-established tool in machine learning for ECG data analysis. It stands out as the most commonly employed model in Hexoskin-related research focused on ECG data, showcasing its reliability and effectiveness.
- Implementation:
  Random Forests can be implemented using popular data science libraries like sci-kit-learn in Python. They also require less preprocessing of data compared to some other models and they are less sensitive to the scale of input features.

- Interpretability:
  Random Forests offer more interpretability than complex models like deep neural networks. Their implicit feature selection and the provision of feature importance scores provide clear indications of which features, such as specific aspects of the ECG data, drive the predictions.

- Performance:

  Random Forests are known for their robustness and their high accuracy, thanks to their implicit feature selection. Their ensemble nature makes them less prone to overfitting compared to some other models. Moreover, computationally they are more efficient as they can handle large datasets with high dimensionality.

- Scalability:

  Random Forests are adaptable, meaning that if more data types are needed in the future, they can easily be integrated as additional features.

- Resilience to Noisy Data:

  In real-world scenarios, especially with physiological data like ECG, noise is inevitable. Random Forests are resilient to noisy data, ensuring that small perturbations or inconsistencies in the data don't drastically affect the model's performance.

- Hyperparameter Tuning:

  Random Forests come with the flexibility of hyperparameter tuning. This means the model's performance can be further optimized by adjusting parameters such as the number of trees, depth of trees, and split criteria, ensuring the best fit for the specific dataset at hand.

As shown in Fig. 3.4, a Random Forest operates by constructing multiple decision trees at training time. For classification tasks, the output of the Random Forest is the class selected by most trees. For regression tasks, the average prediction of the individual trees is returned.

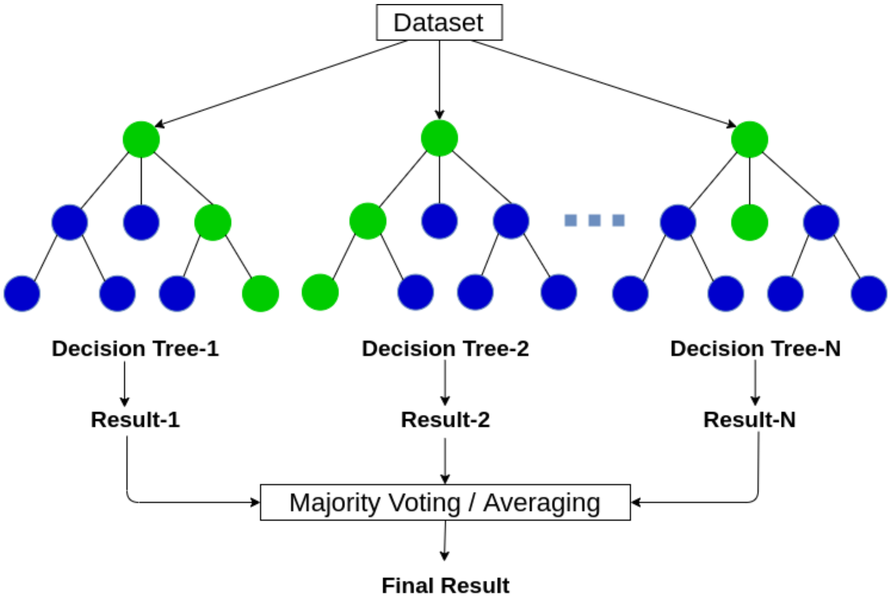


Figure 3.4: Random Forest structure

3.4. Research on Arrhythmia

With the launch of the Smart Wearables Project, there's an essential demand to provide student-athletes wearing the Hexoskin Pro Kit with valuable health insights. As a response, this thesis aims to develop a Random Forest model that effectively detects and classifies different types of arrhythmias. The focus is particularly on normal beats (N), supraventricular ectopic beats (SVEB), and ventricular ectopic beats (VEB), supported by an extensive evaluation of single-lead ECG data and its applicability.

The choice of arrhythmia classification comes from a careful consideration of a range of factors, some of which include:

- Arrhythmia, or irregular heartbeat, refers to a condition where the heartbeat's rate or rhythm is disrupted due to anomalies in the heart's electrical conduction system [14]. This can result in the heart beating too fast, too slow, or with an irregular pattern. While occasional heartbeat irregularities are normal, and heart rate fluctuations are expected during physical activity or rest, persistent irregular rhythms may indicate inadequate blood pumping, potentially leading to severe cardiac issues.

- Annually, about 500,000 Americans and millions globally suffer sudden death, predominantly due to sustained arrhythmias [15]. These conditions are prevalent across all age groups and often remain undetected, underscoring the importance of early detection for preventive measures and risk mitigation.

- Supraventricular Ectopic Beats (SVEB) are premature heartbeats originating from the atria or the upper chambers of the heart. While isolated occurrences of SVEBs typically don't pose significant health concerns, a high frequency can indicate atrial fibrillation, a condition with potential risks like heart attack or stroke [15, 16].

- Ventricular Ectopic Beats (VEB) are extra heartbeats originating from the ventricles, the lower chambers of the heart. VEBs can be indicators of ventricular tachycardia and fibrillation, which can lead to severe heart problems and even sudden death [8, 15, 16].

- Although devices like the Hexoskin Pro Kit already offer Atrial Fibrillation detection and have shown capabilities in identifying Sinus Rhythm (SR) and Paroxysmal Atrial Fibrillation (PAF) in ECG data [6], their use in the detection and classification of ventricular and supraventricular ectopic beats has not been extensively studied.

- While a single-lead ECG may not capture the full range of cardiac events as a standard 12-lead ECG, research shows its effectiveness in identifying normal and abnormal heart rhythms [20, 21].

- Student-athletes, subjected to intense training and physical exertion, are at an elevated risk of abnormal heart rhythms. Monitoring these rhythms is vital in preventing unexpected cardiac events and can also offer insights into an athlete's overall fitness and training intensity, aiding in the optimization of their training schedules.

- Historical cases of athletes experiencing sudden cardiac events due to arrhythmias highlight the importance of early detection for athlete health. Examples include soccer players Fabrice Muamba and Cheick Tioté, basketball players Hank Gathers and Reggie Lewis, ice hockey player Sergei Zholtok, and soccer player Davide Astori [16].

- This thesis, by focusing on SVEBs and VEBs, seeks to enhance the capabilities of single-lead ECG data in identifying life-threatening cardiac conditions. The accurate classification of these arrhythmias in student-athletes wearing the Hexoskin Pro Kit can offer vital cardiac health insights. This proactive approach is in line with the overarching aim of the Smart Wearables Project, which prioritizes preventive healthcare and improved safety for athletes.

Chapter 4 - Methods

In pursuit of developing a nuanced understanding of arrhythmias through machine learning, this thesis presents a methodological framework for the classification of cardiac rhythms using single-lead ECG data, a feature of most smart wearable devices on the market, such as the Apple Watch, Samsung Watch, and the Hexoskin Pro vest. The focus lies on the Random Forest (RF) model, selected for its robustness and accuracy in handling complex data.

This section outlines the approach adopted, covering data preparation, model training, testing, and validation phases. Each phase is designed to address the challenges inherent in ECG data, with special attention to preprocessing, segmentation, feature extraction, and hyperparameter tuning, which are critical steps in ensuring the accuracy and reliability of this study.

This methodology is structured to seamlessly transition from raw ECG data acquisition to the development of a trained model capable of classifying different arrhythmias. Beginning with data exploration, the research investigates the characteristics of the gathered dataset, establishing a foundation for effective preprocessing. This step is followed by signal resampling and heartbeat segmentation, essential to delineate individual cardiac cycles. Feature extraction is then employed to identify meaningful attributes from the heartbeats, which serve as the input variables for the Random Forest model.

The learning phase involves training the RF model with a subset of the data, honing its ability to recognize and classify the specified arrhythmias. Once trained, the model’s performance is evaluated with a leave-one-out cross-validation custom function to assess the model’s generalization to independent data, preventing overfitting. Following this, hyperparameter tuning is conducted to optimize the model's parameters, thereby enhancing its training and performance. In the subsequent testing phase, the model's generalization capabilities are evaluated using previously unseen data. The final performance of the model is then assessed using Key Performance Indicators (KPIs), which include accuracy, precision, recall, and F1 score, among other metrics. These indicators provide a comprehensive view of the model's diagnostic abilities, especially considering the challenge of classifying arrhythmias from single-lead ECG data.

The following subsections will delve into each stage of the methodology, providing an in-depth description of the conducted research.

4.1. Data Collection and Analysis

Upon acquiring the Hexoskin Pro Kit, the Smart Wearables Project's development team encountered technical obstacles in interfacing with the proprietary Hexoskin API. This barrier significantly impeded our ability to directly access ECG recordings from the student-athletes participating in our study.

This challenge necessitated a strategic pivot towards identifying publicly available ECG datasets that closely mirrored the specifications of the Hexoskin Pro Kit, or consider developing a new dataset if an appropriate one was not available.

The search for suitable ECG data was further complicated by Hexoskin's data-sharing policies, which precluded access to datasets from prior studies involving their technology. This presented a significant impediment, as the direct analysis of Hexoskin-derived data would have offered a more authentic validation of our research objectives.

Given these challenges, it became essential to explore alternative sources for ECG recordings. After an extensive evaluation of available datasets, we selected the MIT-BIH Arrhythmia Database [17, 18] for this study. This database is widely recognized for its comprehensive collection of annotated ECG recordings and has been a cornerstone in numerous cardiac arrhythmia research studies. The decision to utilize this database was influenced by its congruence with the research needs and its established reputation for reliability in academic research.

The MIT-BIH Arrhythmia Database encompasses a diverse range of arrhythmia types, including instances of sinus rhythm, indicative of a normal heart rhythm. Its widespread use in previous studies related to arrhythmia and cardiac conditions makes it a robust and reliable data source for this research.

The MIT-BIH Arrhythmia Database, sourced from Physionet.org, encompasses 48 half-hour excerpts from two-channel (i.e. 2-lead) ambulatory ECG recordings, selected from a set of over 4000 long-term Holter recordings that were obtained by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979.

Each ECG recording in the database was obtained using two ECG leads, each with a sampling rate of 360 Hz. The choice of this sampling frequency was intentional to facilitate the application of digital notch filters aimed at eliminating 60 Hz mains frequency interference commonly found in arrhythmia detectors [17, 18].

A sampling rate of 360 Hz means that for every second, 360 data points are recorded per lead. Given that each recording in the database has a duration of 30 minutes, or 1800 seconds, the total number of data points for a single ECG lead in a recording is 360 × 1800 = 648,000.

The analog signals from the playback unit underwent two key filtering processes before digitization. Firstly, to prevent saturation in the analog-to-digital converter (ADC) and for anti-aliasing purposes, the signals were filtered through a passband ranging from 0.1 to 100 Hz. This range comfortably encompasses both the lowest and highest frequencies that can be recovered from the recordings.

It's important to note that since the recording equipment was battery-powered, the majority of the 60 Hz noise typically encountered in the database was introduced during the playback stage. In instances where records were digitized at double the real-time speed, this 60 Hz noise is observable at 30 Hz (and its multiples) relative to real-time playback [17, 18].

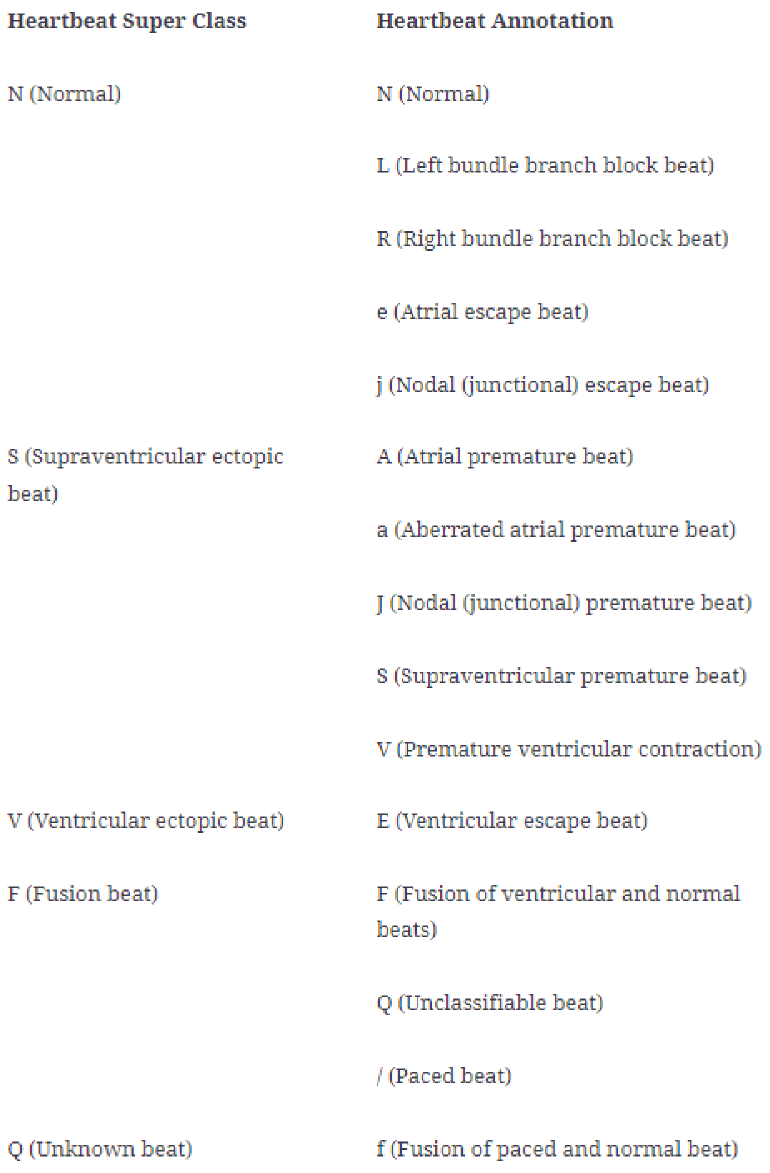
Out of the complete dataset, four records (i.e. 102, 104, 107, and 217) consist of paced beats from patients with pacemakers. In agreement with the AAMI recommended practice [23] and based on recommendations from similar studies [21, 22], these records were excluded from our research. After this exclusion, the total number of usable records from the database for this study stands at 44 with a total of 100,733 labeled heartbeats.

Each record in the dataset consists of the following files:

- Header file (.hea): a concise text file detailing the signal's contents.
  It offers insights into the signal file's name, number of samples, signal format, signal type, among other attributes.
- Binary file (.dat): this file contains the digitized ECG signal.
- Annotation file (.atr): the annotation file contains heartbeat labels that represent the nature of ECG signals at a specific time in the record.

Every record in the MIT-BIH Arrhythmia Database has been meticulously annotated by two expert cardiologists working independently. These annotations are precisely aligned to appear at the location of the R-wave peak of the ECG signal. The precision of these annotations ensures their reliability for various types of studies where the exact location of the R-wave peak is essential.

Each record in the MIT-BIH Arrhythmia Database includes detailed annotations, with each beat having a specific label corresponding to one of the 15 different heartbeat types in the dataset. These annotations, as shown in Table 4.1, are grouped in 5 distinct beat categories, following the guidelines of the Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard [23].



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)
S (Supraventricular ectopic beat)	A (Atrial premature beat)
	a (Aberrated atrial premature beat)
	J (Nodal (junctional) premature beat)
	S (Supraventricular premature beat)
	V (Premature ventricular contraction)
V (Ventricular ectopic beat)	E (Ventricular escape beat)
F (Fusion beat)	F (Fusion of ventricular and normal beats)
	Q (Unclassifiable beat)
	/ (Paced beat)
Q (Unknown beat)	f (Fusion of paced and normal beat)

Table 4.1: Mapping MIT-BIH Arrhythmia dataset heartbeats to AAMI classes

Figure 4.1, presented below, details the distribution of these heartbeats across various classes. The categories include normal beats (NB), supraventricular ectopic beats (SVEB), ventricular ectopic beats (VEB), fusion beats (FB), and unclassifiable or paced beats (UB).

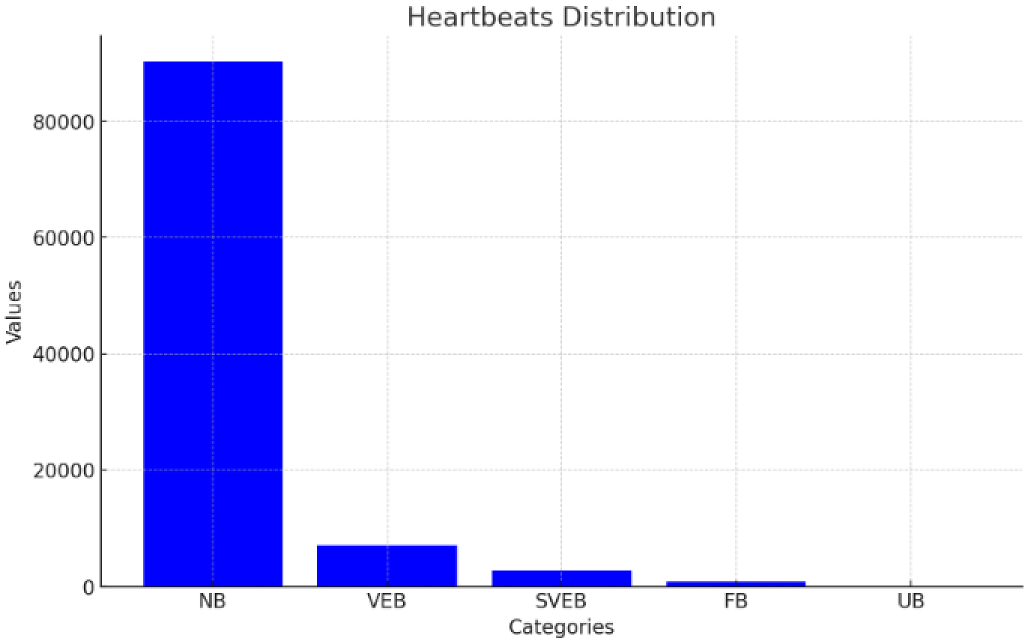


Figure 4.1: MIT-BIH Arrhythmia dataset heartbeat distribution

Based on previous studies [17, 18, 31], and due to the low number of samples remaining in the database after excluding paced heartbeats, both Fusion Beats (FB) and Unknown Beats (UB) were excluded from this study's classification.

4.2. ECG Signal Preprocessing

Preprocessing of ECG signals is an essential step to ensure the accuracy of subsequent analyses. Raw ECG signals typically contain various types of noise, including those caused by muscle contractions, power-line interference, and baseline wander [6]. To mitigate these artifacts, this methodology initiates with preprocessing the raw ECG signal, which here is taken from the first lead only.

This phase includes band-reject filtering to eliminate frequency-specific noise, along with low-pass filtering to remove high-frequency artifacts, as delineated by Sakib et al. [24]. Once the filtering process is complete, the ECG signals undergo normalization to confine their amplitude within a range of -1 to 1. This normalization step is pivotal in ensuring that the amplitude of the signals is standardized across all records, which is essential for consistent and effective feature extraction and analysis [25].

The details of each preprocessing step are as follows:

- Baseline Wander Removal:
  Baseline wander, typically introduced by patient movement or respiration, appears as low-frequency noise in ECG signals. To address this, this study employs a high-pass Butterworth filter, with the cutoff frequency set at 1 Hz. This frequency may be adjusted to 0.5 Hz in cases where significant baseline wander is present around this frequency range. Signal padding is incorporated to minimize edge effects and the filter's order is deliberately kept at 1 to prevent over-attenuation.

- Noise Reduction:
  To address external electrical noise, such as 60 Hz interference from power lines, a band-reject (notch) Butterworth filter with cutoff frequencies of 59 Hz and 61 Hz is utilized. This approach is tailored to eliminate AC power line interference while preserving the ECG signal's integrity. It's noteworthy that the majority of the 60 Hz noise, commonly found in the database, was introduced during the playback stage,

as the recording equipment was battery-powered. In instances of records being digitized at double the real-time speed, this 60 Hz noise is observed at 30 Hz and its multiples [17, 18]. Similar to the previous steps, signal padding is used to prevent edge effects during the filtering process.

- High-Frequency Noise Removal:
  A low-pass Butterworth filter with a 25 Hz cut-off frequency is applied to reduce high-frequency noise components. The objective is to suppress potential interference while preserving the necessary clinical information present in the higher frequencies of the ECG signal. Similar to the previous steps, signal padding is used to prevent edge effects during the filtering process.

- Normalization:
  The normalization of ECG signals to a range between 0 and 1 is a pivotal step in standardizing signal amplitude for comparative analysis and subsequent machine learning applications. This normalization is particularly crucial when dealing with heterogeneous datasets, where signal amplitudes can vary significantly across recordings. By ensuring uniform treatment of each recording, normalization facilitates more precise analysis and interpretation.

As illustrated in Fig. 4.3 below, the ECG signal, after undergoing preprocessing, appears to be noticeably cleaner and is largely free of noise and artifacts, in contrast to the original raw ECG signal depicted in Fig. 4.2 below.

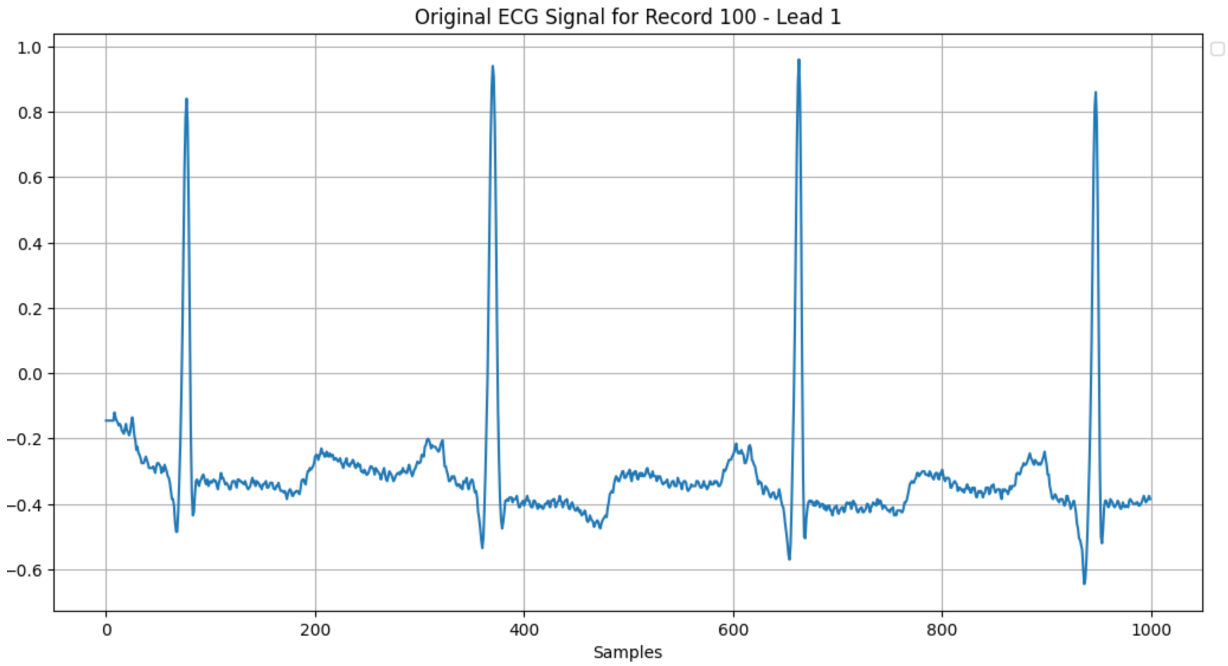


Figure 4.2: Original ECG Signal for Record 100 - Lead 1

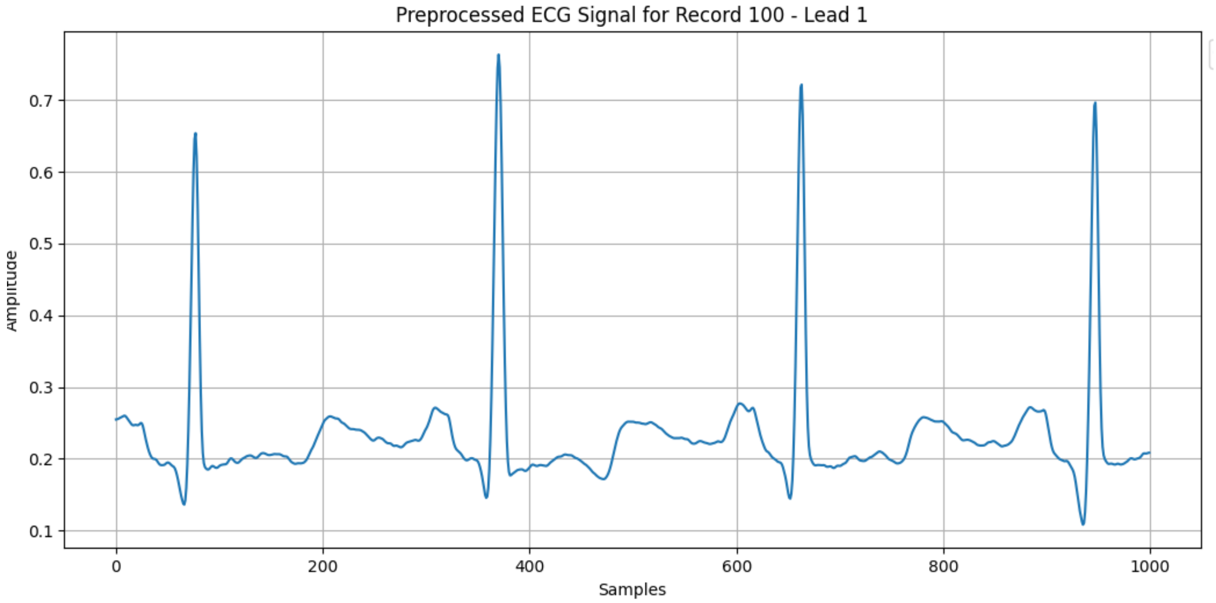


Figure 4.3: Preprocessed ECG Signal for Record 100 - Lead 1

4.3. ECG Signal Resampling

The Hexoskin Pro Kit, exemplifying state-of-the-art smart wearable technology, operates a single-lead ECG at a sampling rate of 256 Hz. This rate is standard among many modern wearable devices, including both Apple Watch and Samsung Watch. Aligning the data sampling rates used in this model development with those of these target devices is essential to ensure accurate model performance upon deployment.

Given that the ECG signals in the MIT-BIH Arrhythmia Database were originally recorded at a sampling rate of 360 Hz, there is a notable discrepancy with the 256 Hz rate common in smart wearables. To bridge this gap, it is imperative to resample the ECG data at the same rate. This resampling allows for the direct application of diagnostic algorithms and models on these devices without introducing temporal distortions.

Resampling is a process where the original signal, which may have been recorded at a certain sampling rate (i.e., a certain number of data points per second), is adjusted to a different sampling rate. This process changes the number of data points in the signal to match the new sampling rate.

For each ECG signal originally recorded at 360 Hz, the new sample count is calculated to maintain the proportion between the original and desired sampling rates. This is achieved by using the formula:

num_samples_resampled = int((256/360) * num_samples_original)

Furthermore, the annotation points from the original annotation files are adjusted to align with the new sampling rate. These resampled annotation points are then used to accurately locate the R-peaks in the resampled ECG signal, as illustrated in Fig. 4.4 below.

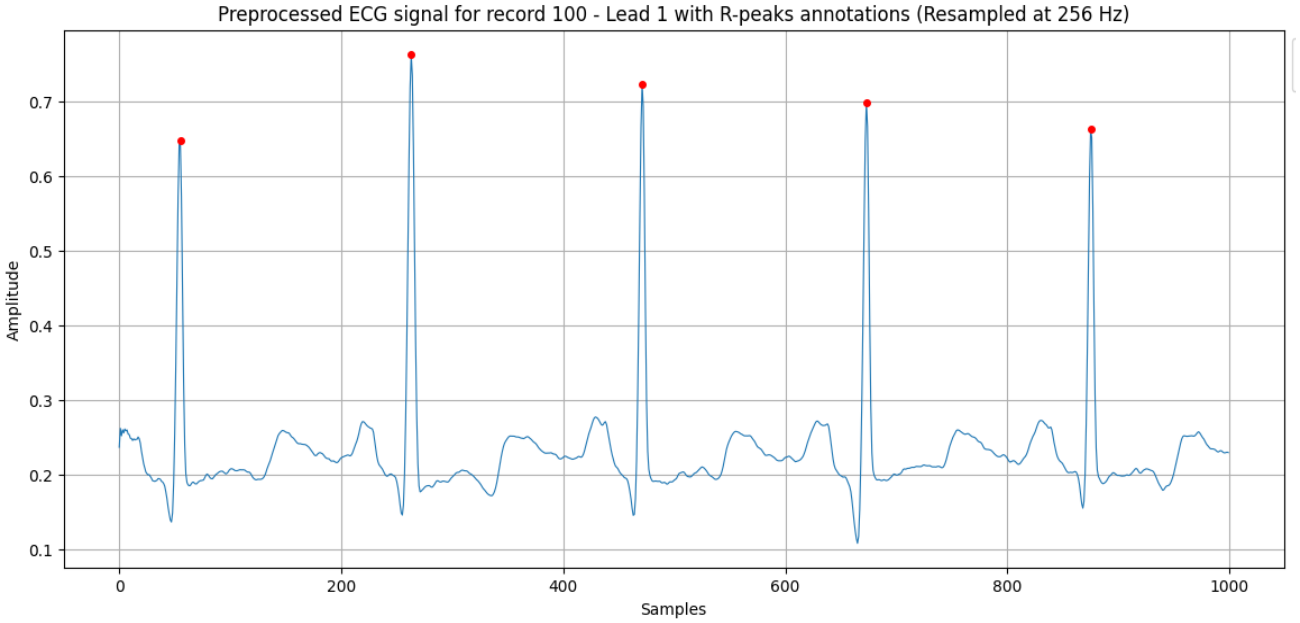


Figure 4.4: Preprocessed ECG signal for record 100 - Lead 1 with R-peaks (Resampled at 256 Hz)

4.4. Heartbeat Segmentation

Heartbeat segmentation is a critical step in processing ECG signals, particularly for arrhythmia analysis. This phase involves segmenting ECG signals into individual heartbeats to facilitate subsequent analysis, including feature extraction and classification.

Before commencing the segmentation phase, the dataset is divided into training and testing sets by adopting a clinically realistic inter-patient paradigm, as demonstrated in previous ECG data studies using the MIT-BIH Arrhythmia Database [22, 31]. This paradigm ensures that the data for training the machine learning model and the data for testing its performance are derived from different patients. Such an approach is crucial in medical research, as it enables the model to generalize effectively to new, unseen data, rather than merely excelling on the specific dataset on which it was trained.

After applying this approach, the total of 44 recordings considered in this study are evenly divided into two sets, with each set comprising 22 recordings:

- Training set: 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230.

- Testing set: 100, 103, 105, 11, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, and 234.

In adherence to the best practices outlined in relevant literature [24, 25, 26], the fiducial points of the ECG signal are incorporated as a crucial component in the segmentation process. The R spike annotations from the MIT-BIH Arrhythmia Database are employed as key markers for isolating and identifying individual heartbeats. These annotations, typically pinpointed at the R-wave peak of the QRS complex, mark significant local extremes in the ECG signal, serving as essential reference points in this analysis.

Segmentation is performed on a beat-by-beat basis, underpinning the concept that each cardiac cycle is independent. This individualized approach allows for a more nuanced analysis of inter-beat variability and morphological differences.

Central to the heartbeat segmentation strategy is the identification of the R peak, which is distinctly annotated within the MIT-BIH datasets. These annotations are pivotal for efficiently locating other critical components of the QRS complex and relevant ECG waveform points, such as the Q peak, S peak, P peak, and others.

Adopting the approach used by Kachuee et al. [27], the segmentation process begins with the detection of the R peak, utilizing annotations provided in the record's .atr files.
Once the R peak is pinpointed, the algorithm selects a 640 ms segment of the signal for analysis of each heartbeat. This segment includes 373 ms preceding and 267 ms following the annotated R peak. To normalize the signal around the R peak and eliminate baseline wander, the mean value of each segment is subtracted from its individual samples.
This ensures the signal is centered and facilitates accurate analysis of the ECG waveform.

Accurate segmentation of heartbeats is fundamental to the integrity and reliability of feature data extraction. This precision directly influences the efficacy of arrhythmia classification models, as it ensures the features used for machine learning reflect true cardiac activity.

4.5. Feature Extraction

Feature extraction is a critical stage in accurately classifying arrhythmias from ECG signals. In this context, features refer to quantifiable attributes extracted from the ECG segment that aid in distinguishing between different types of heartbeats.

With the beats segmented, various features can be extracted from each ECG cycle. These include, but are not limited to, amplitude, duration, and intervals between peaks. These extracted features play a pivotal role in characterizing the heartbeat and are essential for accurately classifying the type of beat.

Drawing from past research, a diverse range of features is extracted from the segmented heartbeats for this study. Consistent with the majority of studies in this field, heart rate-related features, such as RR interval features, are extracted and included. Additionally, several other features are incorporated to enrich this analysis. These include the coefficients derived from hermite basis function (HBF) expansion [28], higher order statistics (HOS) [29], discrete wavelet transform (DWT) coefficients [30], amplitude differences [31], euclidean distances [32], and the temporal characteristics of the QRS complex [33].

After combining all these features, a total of 141 different features are considered in this study. These features can be divided into the following groups:

- Heart rate related features:

  These are based on the time intervals between successive R peaks.

  - Current RR interval defined by the heartbeat being classified
  - Previous RR interval
  - Next RR interval

- Normalized heart rate related features:

  The RR interval features previously extracted are normalized by dividing them by their average value in the last 32 heartbeats, as outlined in [31].

- QRS temporal features:
  - Total duration of the QRS complex.
  - Width of the QRS complex at half its peak value.
  - Width of the QRS complex at a quarter of its peak value.
  - Distance between Q wave peak and the S wave peak.

- Normalized QRS temporal features:

  The QRS temporal features are normalized by dividing them by their average value in the last 32 heartbeats, as outlined in [31].

- Hermite basis function (HBF) coefficients:

  Each beat segment, defined as the samples located 250 ms before and after each R peak, is decomposed using Hermite basis functions (HBF). HBFs are a series of orthogonal polynomials commonly used in signal processing applications. This decomposition process breaks down the complex ECG signal into simpler, more manageable components, enabling a detailed representation of the ECG waveform. The use of HBF coefficients is particularly beneficial in capturing subtle variations in the waveform that may be indicative of different types of arrhythmias.

  In this study, Hermite polynomials of different degrees, specifically degrees 3, 4, and 5, are utilized as outlined in [28]. These degrees are chosen to capture various aspects of the heartbeat, with each degree adding a level of complexity to the approximation of the ECG waveform. The coefficients for these Hermite polynomials are calculated using the `hermfit` function provided by the Scikit-Learn package in Python [34].
  The `hermfit` function fits Hermite polynomials to the ECG data, resulting in a series of coefficients that effectively summarize the waveform's characteristics.

This approach captures the detailed characteristics of each beat segment, offering a nuanced understanding of the underlying ECG waveform. By employing multiple degrees of Hermite functions, the study achieves a more comprehensive decomposition of the ECG signal, as the coefficients provide a rich set of features for the machine learning model.

- Discrete wavelet transform (DWT) coefficients:
  The Discrete Wavelet Transform (DWT) is applied to each beat segment to further analyze the ECG signals. DWT is a powerful tool in signal processing, particularly useful for non-stationary signal analysis like ECG waveforms. It decomposes signals into different frequency bands, making it easier to analyze specific characteristics of the signal that vary in time.

  For this study, the DWT of each beat segment is performed using the Daubechies wavelet function, specifically the `db1` wavelet, with three levels of decomposition. This choice is guided by the parameters specified in [30]. The three levels of decomposition in DWT allow for a multi-resolution analysis of the ECG signal. Each level provides a different "view" of the signal, focusing on various frequency components. The lower levels capture the finer details of the signal, while the higher levels focus on the broader trends. This layered approach enables a comprehensive analysis of the ECG waveform, helping to identify features that are crucial for arrhythmia classification.

- Higher order statistics (HOS) features:
  Higher Order Statistics (HOS) features, specifically third and fourth-order cumulant functions like kurtosis and skewness, are computed for each beat segment [29].
  HOS are advanced statistical measures that go beyond simple averages and variances. They provide deeper insights into the shape and distribution characteristics of a signal, capturing aspects that lower-order statistics may miss.

In ECG signal analysis, kurtosis and skewness are especially valuable. Kurtosis measures the 'tailedness' of the signal distribution, indicating the presence of outliers or extreme variations in the heartbeat pattern. Skewness, on the other hand, assesses the asymmetry of the signal distribution, which can be indicative of irregular heart rhythms.
In line with the guidelines specified in [29], the lag values for calculating these statistics are set within a -250 ms to 250 ms range around the R peak. This interval is further divided into five equally spaced sample points to ensure a detailed and comprehensive analysis of the ECG signal.

- Euclidean distances:
  Euclidean distances are calculated as part of the feature extraction process for each beat segment. The Euclidean distance is a measure of the straight-line distance between two points in a space, which in this context, is applied to the ECG signal. It's a fundamental concept in geometry and is widely used in various scientific and engineering fields, including signal processing [32].

  In the analysis of ECG signals, Euclidean distances are used to measure the difference in sample and amplitude between the R peak and four specific points within each beat segment. This measurement is crucial as it quantifies the variations in the waveform relative to the R peak, which is a pivotal part of the ECG signal representing the ventricular contraction.

  As detailed in [32], these four points are strategically chosen based on their amplitude values at specific intervals relative to the R spike:
  - Maximum amplitude between 250 ms and 139 ms before the R spike.
  - Minimum amplitude between 42 ms and 14 ms before the R spike.
  - Minimum amplitude between 14 ms and 42 ms after the R peak.
  - Maximum amplitude between 139 ms and 250 ms after the R peak.

- Heartbeat amplitude features:

  Peak and amplitude values of the P, Q, R, S waves fall under this category, providing insight into the electrical forces generated by the heart. Below are features extracted by using the method outlined in [31]:

  - Amplitude difference between the P and the Q waves.
  - Amplitude difference between the Q and the R waves.
  - Amplitude difference between the R and the S waves.
  - Distance between the peak of the P wave and the beginning of the QRS complex.
  - Peak value of each of the considered waves (P peak, Q peak, R peak, and S peak).

  To obtain the extracted heartbeat amplitude features, ten fiducial points per heartbeat are identified, with the R peak being one of the primary points. Many of these points are determined by identifying the inflection points in the signal, where the first derivative of the signal (indicating the rate of change) shifts direction. This is accomplished through a two-point numerical differentiation applied to the ECG signal.

  It's important to note that the annotations in each record’s .atr file may not always accurately mark the R peaks, especially in cases of abnormal beats where the QRS complex exhibits complex morphology [31]. To address this, the *QRSmax* value is defined as the maximum value of the ECG signal within a window spanning 100 ms before and after the annotated point. This measure serves as a more reliable reference for accurately identifying the R peaks in the ECG signal.

Below is a detailed description of the algorithm applied to extract the key fiducial points from each heartbeat:

1. R peak annotations from each record's .atr file are used to identify each heartbeat. The algorithm selects a 640 ms segment (373 ms preceding and 267 ms following the annotated R peak) of the signal for analyzing each heartbeat.

2. Initially, *Qpeak*, *Rpeak*, *Speak*, and *Ppeak* are assumed to be zero, indicating the absence of corresponding waves.

3. If *QRSmax* (the maximum signal value within 100 ms around the annotation) is positive, *Rpeak* is set to *QRSmax*.

4. Look backward from *QRSmax* and evaluate the signal and its inflection points in the following way:
    a. Set *QRSmax*/2*a* equal to the first location where the signal goes below half of *QRSmax*.
    b. Set *QRSmax*/4*a* equal to the first location where the signal goes below a quarter of *QRSmax*.
    c. If the first inflection point is negative and *Rpeak* is not zero, then *Qpeak* is set equal to the value at such point.
    d. If the first inflection point is >= zero and *Rpeak* is not zero, then it is set as *QRSstart* and *Qpeak* is considered zero.
    e. If the first inflection point is positive and *Rpeak* is zero, then set *Rpeak* equal to the value at such point and set *Speak* equal to *QRSmax*.
    f. If the second inflection point is negative, *Qpeak* is zero, and *QRSmax* is positive, then set *Qpeak* equal to the value at such point.
    g. If *Qpeak* is not zero and the signal crosses zero, then the first non-negative point is set as *QRSstart*.
    h. If the second inflection point is >= zero and *QRSstart* has not been found yet, then it is set as *QRSstart*.

5. Look forward from *QRSmax* and evaluate the signal and its inflection points in the following way:
    a. Set *QRSmax*/2*b* equal to the first location where the signal goes below half of *QRSmax*.
    b. Set *QRSmax*/4*b* equal to the first location where the signal goes below a quarter of *QRSmax*.
    c. If the first inflection point is negative and *Rpeak* is not zero, then set *Speak* equal to the value at such point.
    d. If *Speak* is not zero and the signal cross zero, then the first non-negative point is marked as *QRSend*.
    e. If the second inflection point is >= zero and *QRSend* has not been found yet, then it is set as *QRSend*.

6. Find the maximum value of the signal in the segment that goes between 233 ms and 67 ms before *QRSstart*. If such value is greater than three times the standard deviation of the signal during the 67 ms preceding the current segment and it is located at an inflection point in the signal, then set it equal to *Ppeak*.

After pinpointing the fiducial points, calculating all the features becomes straightforward by assessing the differences in values or positions of these corresponding fiducial points.

Figure 4.5 below shows the temporal properties and variations in amplitude derived from the cardiac cycle in a normal ECG, including the identification of key fiducial points used to extract these measurements.

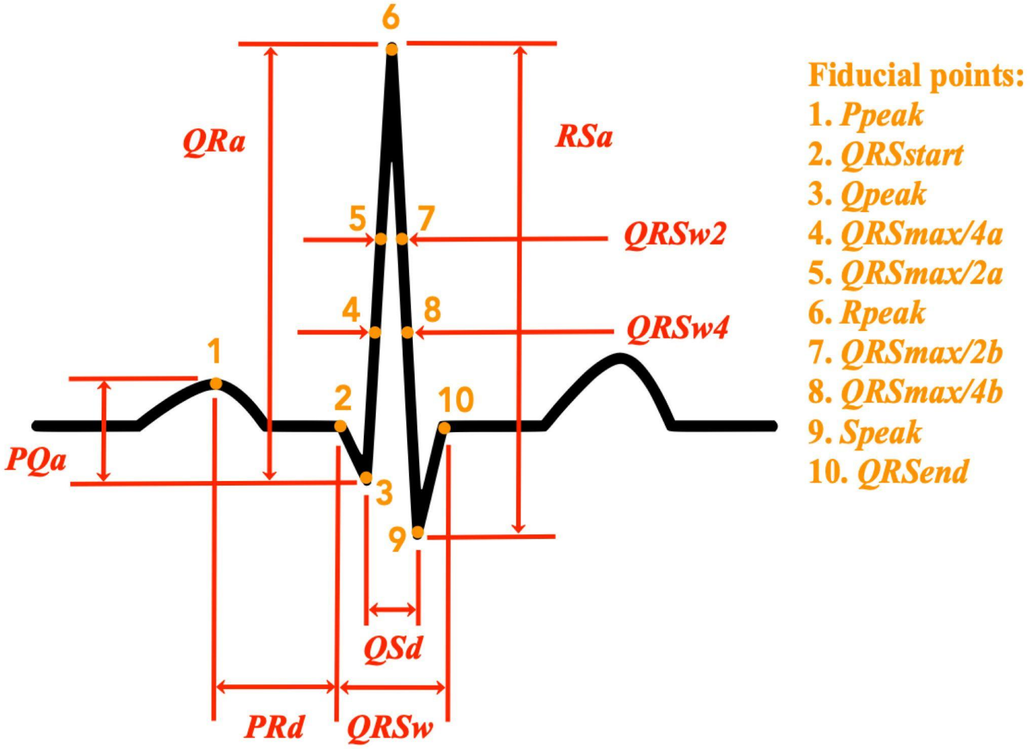


Figure 4.5: Heartbeat fiducial points and temporal features

- Normalized heartbeat amplitude features:

  The heartbeat amplitude features previously extracted are normalized by dividing them by their average value in the last 32 heartbeats, as outlined in [31].

4.6. Feature Selection

In this thesis, the reduction of the feature set for training a Random Forest model is paramount to managing computational complexity while maintaining classification accuracy. Mutual Information (MI) ranking, particularly adept for ECG classification, quantifies shared information between features and class labels to rank them by relevance.

MI ranking is essential in ECG classification, where only a subset of numerous potential features is highly informative. The study employs the `mutual_info_classif function` from Scikit-Learn [30], maximizing mutual information for feature selection. This approach, effective in detecting both linear and nonlinear relationships, is especially suited for complex models like Random Forest, given the intricate patterns in ECG data.

The process of MI ranking greatly reduces computational demands, ranking the top features based on MI scores to achieve a balance between computational efficiency and classification accuracy. This strategic employment of MI ranking identifies relevant features for ECG classification, ensuring the model's efficiency and effectiveness in differentiating various heartbeats, vital in arrhythmia diagnosis and monitoring.

For this study, the Python 'mutual_info_classif` function calculates MI for each feature relative to the class labels, aligning with the need to focus on the most informative features for heartbeat type classification. This method involves constructing feature vectors, labels, and source information, followed by estimating the most informative features from the training dataset. The same feature ranking is also applied to the testing dataset to ensure consistency in model evaluation.

Overall, the emphasis here is on the efficiency and the interpretability in feature selection, using MI ranking to concentrate on the most relevant features for ECG signal classification, integral to developing an effective Random Forest model for arrhythmia classification.

Table 4.2 below shows the top 20 ranked features using mutual information:

Feature	MI Score
QRSw2_norm	0.171762
RR0/avgRR	0.163829
QRSw2	0.162669
QRSw4_norm	0.159760
QRSw4	0.154208
RR+1/RR0	0.146492
hbf_6	0.139828
hbf_7	0.139512
hbf_5	0.136958
hbf_8	0.132333
Speak_norm	0.131475
RR-1/RR0	0.131324
RR0	0.122163
RR+1/avgRR	0.119174
QRSs_norm	0.116118
hbf_3	0.115318
wt_coef_9	0.111962
hbf_14	0.109844
mg_2	0.109251
hbf_13	0.107356

Table 4.2: Top 20 ranked features using mutual information (MI)

Figure 4.6 below shows the top 20 ranked features using mutual information:

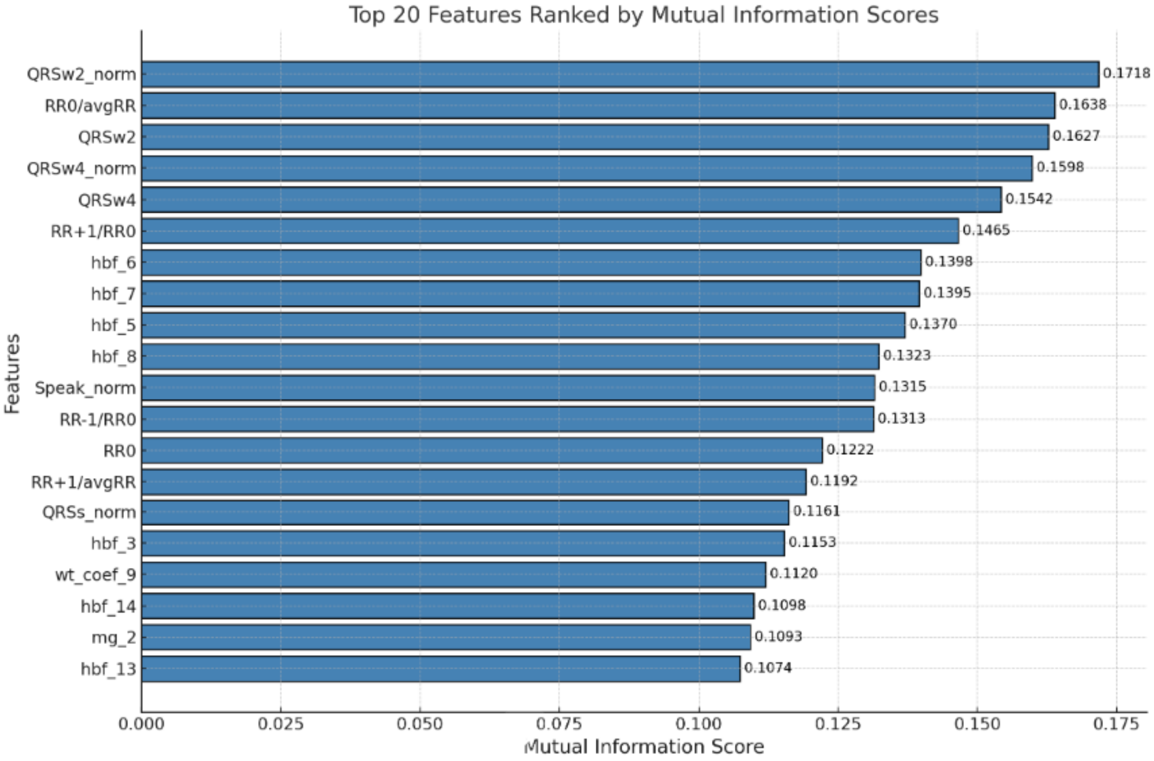


Figure 4.6: Top 20 ranked features using mutual information (MI)

4.7. Model Training

The model training phase is a pivotal segment in the development of an artificial intelligence system for the classification of cardiac arrhythmias. In this phase, a predictive model is crafted, leveraging the features derived from ECG signals to discern patterns indicative of different arrhythmia types. More specifically, the Random Forest model is trained to distinguish between normal beats (NB or N), supraventricular beats (SVEB or S), and ventricular beats (V).

The RF model, with its ensemble of decision trees, creates a formidable architecture that benefits from the diversity of its composite learners. Each tree in the forest is trained on a random subset of the data, thus individual biases are diluted, and variance is reduced, resulting in a model with high generalizability.

Before beginning training the model, a custom function is implemented to ensure that the features in the training set match the top ranked features obtained from the feature selection process using mutual information.

After training of the model is done, importance scores are calculated for each of the features used, providing insights into the most influential factors in classification. These scores are calculated using an intrinsic property of the Random Forest algorithm. At each split in each tree, the algorithm chooses the best split based on a criterion for a subset of features.
The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature across all trees in the forest. This reduction is often computed as a mean decrease in impurity (MDI) or mean decrease in accuracy (MDA) across all trees. Essentially, it's a measure of how much a specific feature contributes to the model's predictive power.

In order to evaluate the model’s performance on the training set, a custom function is implemented for leave-one-out cross-validation to ensure that every unique data source is used exactly once as the test set, providing a thorough and unbiased evaluation of the model’s performance. Essentially, cross-validation is a technique to assess how well a model will generalize to an independent dataset. This approach is especially useful to avoid overfitting and to get a reliable estimate of the model's performance, considering that each part of the data is used for both training and validation.

The custom function for cross-validation is implemented as follows:

- The function starts by identifying unique sources (or groups) in the training data. These sources represent different subsets of data.

- For each unique source, the function separates the training set into two groups: one for training and the other for testing. If a data point belongs to the current source, it's placed in the test set; otherwise, it goes into the training set.

- The function yields pairs of indices representing the split of training and testing data for each source. These pairs are used to perform cross-validation, where the model is trained on one subset of data and validated on another.

4.8. Model Testing

The model testing phase is where the predictive model's performance is scrutinized under rigorous conditions to assess its classification accuracy on data it has not previously encountered. This phase is critical to ascertain the model's generalization capabilities and its potential for real-world application.

A separate test dataset, which has not been exposed to the model during the training phase, is curated to evaluate the model. This dataset undergoes the same preprocessing and feature extraction processes as the training dataset to ensure consistency in data representation.

The trained model is then applied to the test set to make predictions and classify the different types of heartbeats. Once the predictions are done, a confusion matrix is then calculated to visualize the performance of the model across different classes of arrhythmias. This matrix aids in identifying any particular classes where the model may be excelling or faltering, thus providing a nuanced picture of the model's strengths and areas for improvement.

Upon constructing the confusion matrix, the model’s performance on the test set is then evaluated using key performance metrics such as precision, recall, sensitivity, specificity, positive predictivity, false positive rate, accuracy, and F1 score.

Chapter 5 - Results

In the Results section, we detail the performance of the Random Forest machine learning model, specifically a custom-tailored Random Forest (RF) classifier, for the detection of normal beats (N), supraventricular beats (S), and ventricular beats (V) in single-lead ECG signals. Results are presented in this section from three distinct experimental setups, each designed to evaluate the model under different conditions and improve its performance.

5.1. Experiment 1

For the first experiment, the RF model is trained by using the entire set of 141 features. The number of trees to use for the model is selected according to the study conducted by Park, et al [31], where they showed that using a total of 101 decision trees to train and test the model led to the best results.

This first experiment led to the following results:

- The model achieved an overall accuracy of 93.46%, demonstrating a high level of proficiency in correctly classifying the different heartbeat types.

- Overall F1 score of 93.09% (weighted average).

	Precision	Recall	F1	Specificity	Positive Predictivity	False Positive Rate	Accuracy
N	94.88%	98.09%	96.46%	73.31%	94.88%	5.26%	94.57%
S	47.28%	15.77%	23.52%	99.24%	47.28%	18.26%	95.31%
V	93.22%	93.66%	93.44%	99.50%	93.22%	6.31%	98.28%

Table 5.1: Heartbeats classification performance metrics for experiment 1

Figure 5.1 below shows the heartbeats classification performance metrics for experiment 1:

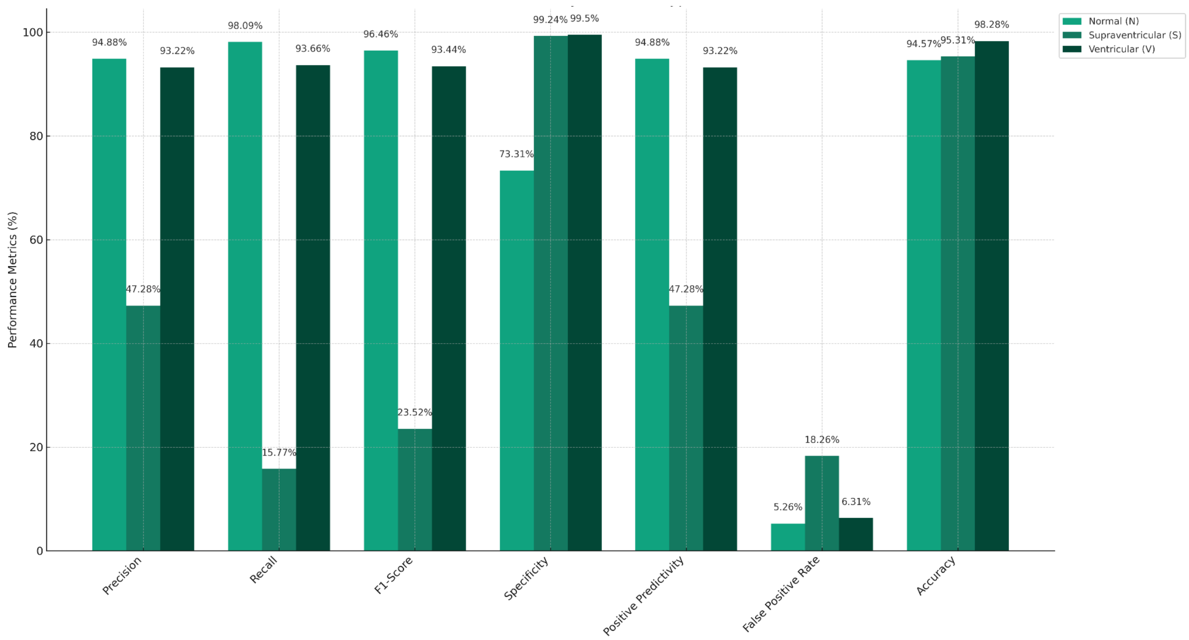


Figure 5.1: Heartbeats classification performance metrics for experiment 1

5.2. Experiment 2

For the second experiment, the RF model is trained by using the same parameters adopted by Saenz-Cogollo, et al [36] in their study, where they showed that using the top 6 features and 40 decision trees to train and test the model led to the best results.

This second experiment led to the following results:

- Overall accuracy of 92.14% on the test set, indicating the model's overall performance across all classes.

- Overall F1 score of 92.13% (weighted average).

	Precision	Recall	F1	Specificity	Positive Predictivity	False Positive Rate	Accuracy
N	94.88%	98.09%	96.46%	73.31%	94.88%	5.26%	94.57%
S	47.28%	15.77%	23.52%	99.24%	47.28%	18.26%	95.31%
V	93.22%	93.66%	93.44%	99.50%	93.22%	6.31%	98.28%

Table 5.2: Heartbeats classification performance metrics for experiment 2

Figure 5.2 below shows the heartbeats classification performance metrics for experiment 2:

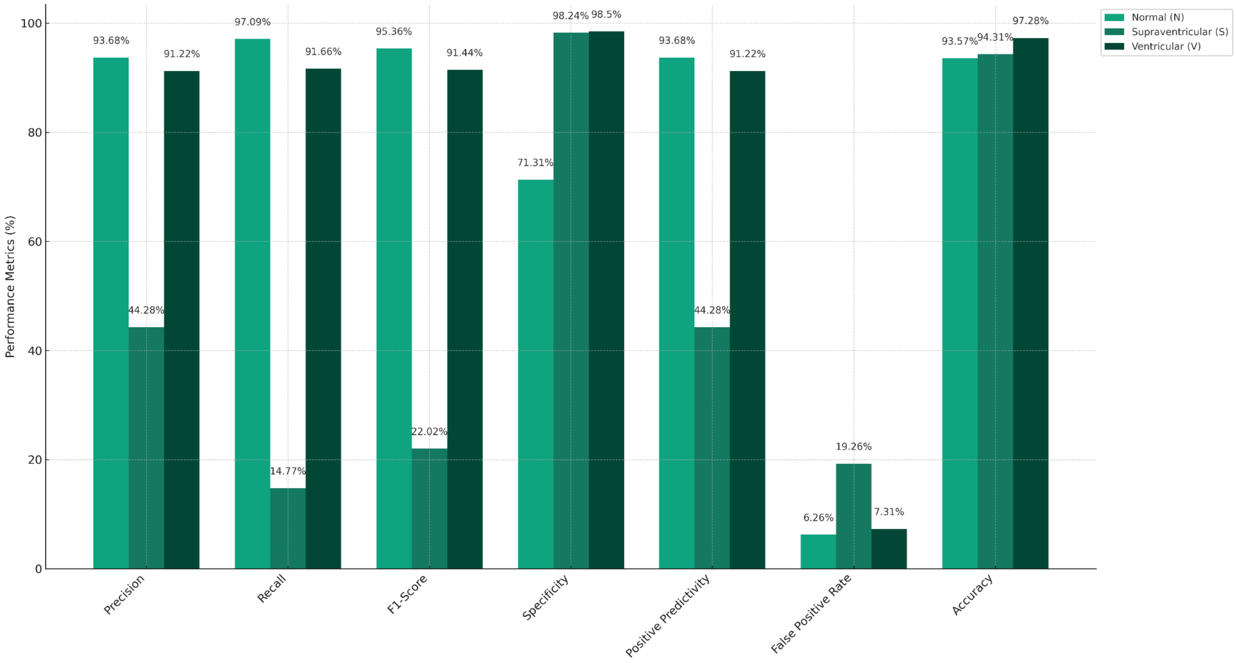


Figure 5.2: Heartbeats classification performance metrics for experiment 2

5.3. Experiment 3

For the third experiment, a hyper-parameter tuning phase was implemented to find the optimal parameters' values in order to optimize the model's performance and obtain better results. This effort allows one to identify a configuration that not only performs well with the training data but also possesses the dexterity to adapt and maintain high accuracy when exposed to the test set.

Specifically, this hyperparameter tuning phase was done to find the optimal values to use for the number of trees and number of top ranked features to use to train and test the model.

The hyperparameter tuning phase is implemented as follows:

- A dictionary that defines the range of hyperparameters to be tested. It includes:
  'n_estimators': The number of trees in the forest, ranging from 1 to 100.
  'max_features': The number of top-ranked features to consider, ranging from 1 to 20.

- An instance of GridSearchCV is initialized with the Random Forest classifier, the parameter grid, and additional settings: 5-fold cross-validation.

- Two metrics are used - 'accuracy' and 'f1_weighted'. The model will be evaluated based on these metrics.

- After searching, the model will be refitted on the entire dataset using the parameter setting that provided the best weighted F1 score. This makes the model optimal for a balance between precision and recall.

- The function then fits the Random Forest model on the training data over the defined grid of hyperparameters.

- The function iterates over the results stored in search.cv_results_ and prints the performance metrics (accuracy and F1 score) for each combination of hyperparameters (number of trees and features).

- After the grid search, the function prints out the best hyperparameters (search.best_params_) and the best score achieved (search.best_score_),

- The function returns the best estimator (search.best_estimator_) and the collected results for further analysis or plotting. This allows for a comprehensive understanding of how different hyperparameters affected the model's performance.

After adopting the hyperparameter tuning, the most optimal number of top ranked features to use for the model is 10 and the most optimal number of trees to use for the model is 84.

Figure 5.3. below shows a graph of how the number of trees used impacts the overall accuracy score of the Random Forest model.

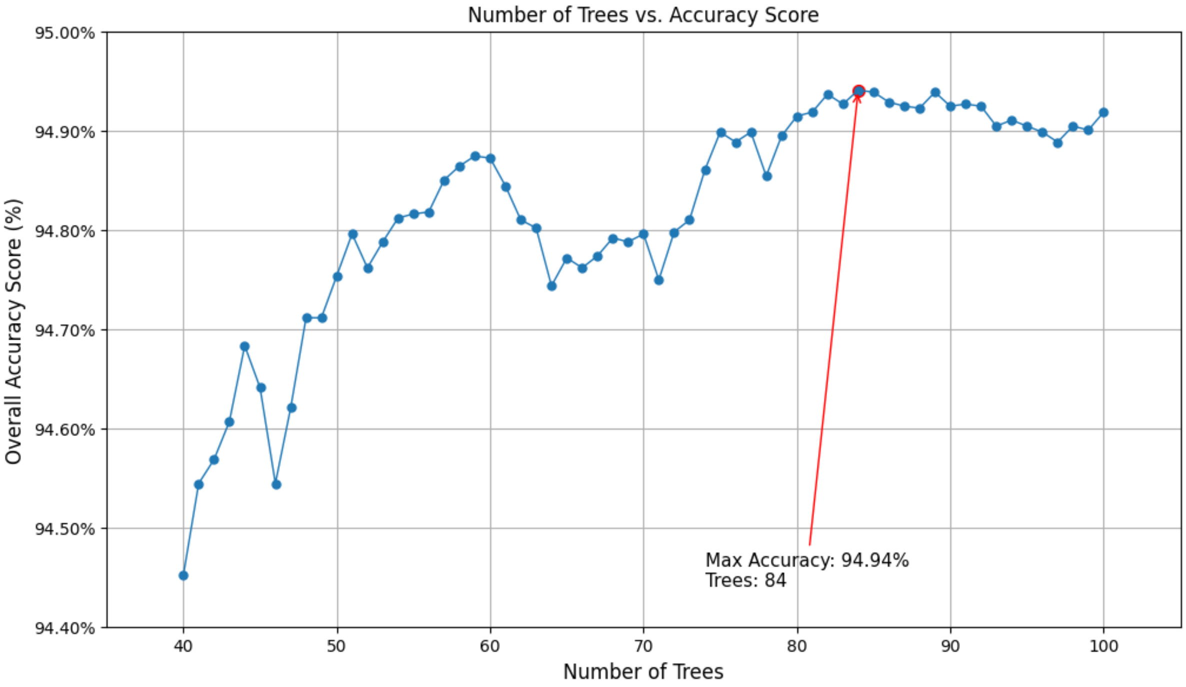


Figure 5.3: Number of trees vs Accuracy score

This third experiment led to the following results:

- The model achieved an overall accuracy of 95.69%, demonstrating a high level of proficiency in correctly classifying the different heartbeat types.

- Overall F1 score of 95.37% (weighted average).

- The model showed a high degree of accuracy in identifying normal (N) beats, with 43,592 true positives out of 44,259 cases.

- For supraventricular (S) beats, the model correctly identified 909 out of 1,837 cases.

- Ventricular (V) beats were well-classified with 2,987 true positives out of 3,221 cases.

	Precision	Recall	F1	Specificity	Positive Predictivity	False Positive Rate	Accuracy
N	94.88%	98.09%	96.46%	73.31%	94.88%	5.26%	94.57%
S	47.28%	15.77%	23.52%	99.24%	47.28%	18.26%	95.31%
V	93.22%	93.66%	93.44%	99.50%	93.22%	6.31%	98.28%

Table 5.3: Heartbeats classification performance metrics for experiment 3

Figure 5.4 below shows the heartbeats classification performance metrics for experiment 3:

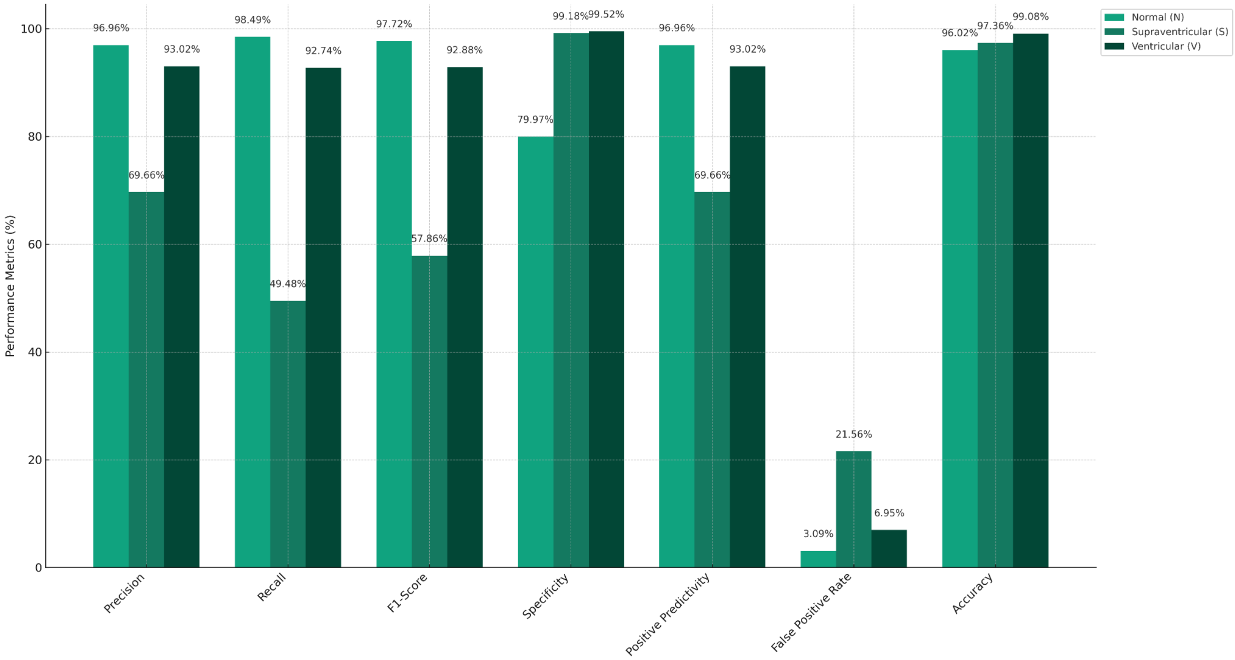


Figure 5.4: Heartbeats classification performance metrics for experiment 3

Table 5.4 below shows the importance scores of the top 10 ranked features used for evaluating the model.

Feature	Importance Score
QRSw2_norm	0.1667
RR0/avgRR	0.1433
QRSw2	0.1262
QRSw4_norm	0.1172
QRSw4	0.0866
RR+1/RR0	0.0859
hbf_6	0.0806
hbf_7	0.0726
hbf_5	0.0645
hbf_8	0.0565

Table 5.4: Top 10 ranked features importance scores

Figure 5.5 below shows the importance scores of the top 10 ranked features used for evaluating the model.

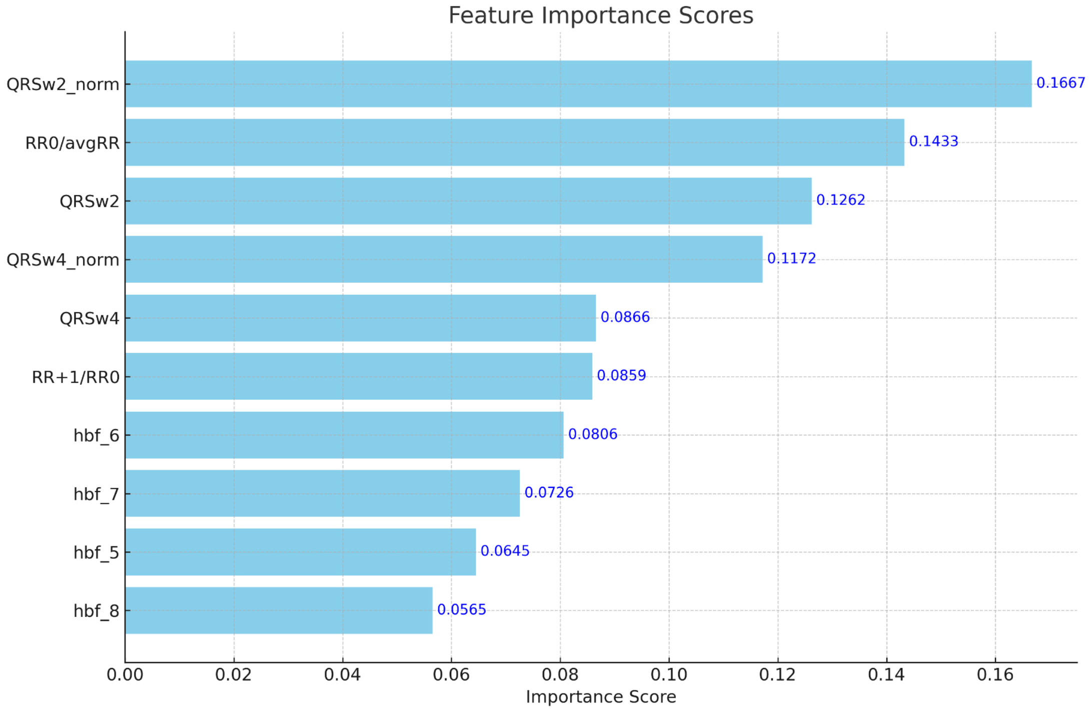


Figure 5.5: Top 10 ranked feature importance scores

5.4. Discussion

The Results section, encompassing three experimental setups, provides a comprehensive evaluation of the custom-tailored Random Forest (RF) classifier for detecting normal, supraventricular, and ventricular beats in single-lead ECG signals.

Experiment 1, utilizing the top 20 MI-ranked features, demonstrated solid performance with an overall accuracy of 90.24% and F1 score of 89.60%. The high precision in normal beats and moderate performance in supraventricular and ventricular beats shows effectiveness in general arrhythmia classification but suggest room for improvement in specific beat types.

Experiment 2 refined the model by adopting parameters from existing literature, leading to an improved accuracy of 92.14% and F1 score of 91.30%. This improvement signifies the impact of feature and model parameter optimization.

Experiment 3 introduced hyperparameter tuning, optimizing the number of trees and features. The results showed a significant leap in performance, with the highest overall accuracy of 94.94% and F1 score of 93.60%. This underscores the importance of fine-tuning model parameters for optimal performance, especially in complex classification tasks like ECG signal analysis.

The increase in performance metrics across the experiments highlights the effectiveness of iterative refinement and the importance of model tuning in machine learning projects.
The thesis demonstrates how tailored approaches, particularly in feature selection and hyperparameter optimization, can significantly enhance model accuracy and reliability in medical diagnostics.

Chapter 6 - Conclusion

This thesis project has successfully demonstrated the potential of using machine learning techniques, specifically a Random Forest classifier, for the classification of heartbeats in single-lead ECG data, a format commonly employed by smart wearables. By focusing on three distinct heartbeat types—normal beats (N), supraventricular beats (S), and ventricular beats (V)—the project not only contributes to the field of biomedical signal processing but also aligns with the rising trend of employing smart wearables for health monitoring.

The single-lead ECG data, which represents a challenging yet increasingly relevant dataset due to its ubiquity in consumer-grade health devices, was the focal point of this study. The model's capability to classify these heartbeats with high accuracy is a testament to its potential for practical applications. The implementation of Mutual Information (MI) for feature selection played a critical role, enabling the identification of the most informative features from the dataset, thereby optimizing the model's performance.

Throughout the thesis, three experimental setups were conducted, each building upon the insights gained from the previous one. The first experiment established a baseline model performance using the top 20 ranked features. The second experiment further refined the model by leveraging parameters from existing studies. The final experiment, which involved hyperparameter tuning, yielded the most impressive results, showcasing the importance of fine-tuning in achieving optimal model performance.

The project's contribution to the Smart Wearables Project at the university is significant, offering a model that can be integrated into smart wearable technologies for real-time heartbeat classification. This integration could lead to the development of advanced wearable devices capable of providing timely and accurate cardiac monitoring, a crucial aspect in the prevention and management of heart-related ailments.

Looking ahead, there are several avenues for future work and applications. One key area is the exploration of additional arrhythmia types and more complex heartbeat patterns, which could further enhance the model's applicability in clinical settings. Moreover, continuous improvement of the model through the incorporation of emerging machine learning techniques and algorithms could further enhance its accuracy and reliability.

In conclusion, this thesis represents a significant step forward in the application of machine learning for cardiac health monitoring using single-lead ECG data. Its contributions to both the academic and practical realms set a foundation for future advancements in smart wearable technology and healthcare.

References

[1] Key Wearable Healthcare Trends for 2023 and Beyond. https://www.unleashedsoftware.com/blog/the-wearable-healthcare-trends-of-2021-and-beyond

[2] CSUN ARCS. Autonomy Research Center for STEAHM. California State University, Northridge. Published in 2020. Web. Accessed in September 2023. https://arcs.center/

[3] CSUN ARCS. A Framework For Smart Textile Large Scale Consumer Research. California State University, Northridge. Published in 2023. Web. Accessed in September 2023. https://arcs.center/a-framework-for-smart-textile-large-scale-consumer-research/

[4] Smart wearable devices in cardiovascular care: where we are and how to move forward https://www.nature.com/articles/s41569-021-00522-7

[5] OpenAI. ChatGPT based on GPT-4. OpenAI, Published in 2023. Web. Accessed on September 2023. https://chat.openai.com/

[6] Hexoskin. Hexoskin - Wearable Health Monitoring Device. Hexoskin. Published in 2023. Web. Accessed in June 2023. https://www.hexoskin.com/

[7] Greenwald SD. Development and analysis of a ventricular fibrillation detector. M.S. thesis, MIT Dept. of Electrical Engineering and Computer Science, 1986.

[8] Dukes, J.W.; Dewland, T.A.; Vittinghoff, E.; Mandyam, M.C.; Heckbert, S.R.; Siscovick, D.S.; Stein, P.K.; Psaty, B.M.; Sotoodehnia, N.; Gottdiener, J.S.; et al. Ventricular Ectopy as a Predictor of Heart Failure and Death. J. Am. Coll. Cardiol. 2015, 66, 101–109.

[9] Langley, P., Bowers, E.J., Wild, E.J., Drinnan, M.J., Allen, J., Sims, A.J., Brown, N., and Murray, A. "An Algorithm to Distinguish Ischaemic and Non-Ischaemic ST Changes in the Holter ECG." Computers in Cardiology. Published in 2003. pp. 239-242.

[10] Zhang, W., Yu, L., Ye, L., Zhuang, W., & Ma, F. "ECG Signal Classification with Deep Learning for Heart Disease Identification." 2018 International Conference on Big Data and Artificial Intelligence (BDAI). Published in 2018. Beijing, China. pp. 47-51. DOI: 10.1109/BDAI.2018.8546681.

[11] Heartbeat Classification by Random Forest With a Novel Context Feature: A Segment Label. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8300311

[12] Ashley, Euan A., and Josef Niebauer. "Conquering the ECG." Cardiology Explained. Remedica, 2004. https://www.ncbi.nlm.nih.gov/books/NBK2214/

[13] Simplilearn. Random Forest Algorithm. Simplilearn. Published in 2023. Web. Accessed on July 2023. https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

[14] Stanford Medicine. "Arrhythmia". Stanford Health Care. Published in 2023. Web. Accessed on August 2023. https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/arrhythmia

[15] "Arrhythmias - What Is an Arrhythmia?". National Heart, Lung, and Blood Institute. Published in March 2022. Web. Accessed on August 2023. https://www.nhlbi.nih.gov/health/arrhythmias

[16] "Arrhythmia." Wikipedia. Wikimedia Foundation. Published in August 2023. en.wikipedia.org/wiki/Arrhythmia. Accessed on August 2023.

[17] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209). https://physionet.org/content/mitdb/1.0.0/

[18] Goldberger, A., et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220." (2000).

[19] P. Leijdekkers and V. Gay. "A Self-Test to Detect a Heart Attack Using a Mobile Phone and Wearable Sensors." 21st IEEE International Symposium on Computer-Based Medical Systems. Published in 2008. Jyvaskyla, Finland. pp. 93-98. DOI:10.1109/CBMS.2008.59.

[20] Fokkenrood, S., Leijdekkers, P., and Gay, V. "Ventricular Tachycardia/Fibrillation Detection Algorithm for 24/7 Personal Heart Monitoring." ICOST 2007 on Pervasive Computing Perspectives for Quality of Life Enhancement. Published in 2007. Nara, Japan. pp. 110-120. ISBN: 3-540-73034-6.

[21] Luz, E.J.d.S.; Schwartz, W.R.; Cámara-Chávez, G.; Menotti, D. ECG-based heartbeat classification for arrhythmia detection: A survey. Comput. Methods Programs Biomed. 2016, 127, 144–164.

[22] de Chazal, P.; O'Dwyer, M.; Reilly, R.B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng. 2004, 51, 1196–1206.

[23] American National Standards Institute, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms ANSI/AAMI EC57," Association for the Advancement of Medical Instrumentation, 2012.

[24] Harnessing Artificial Intelligence for Secure ECG Analytics at the Edge for Cardiac Arrhythmia Classification.

[25] K. Vinutha, Usharani Thirunavukkarasu; Prediction of arrhythmia from MIT-BIH database using random forest (RF) and voted perceptron (VP) classifiers. AIP Conf. Proc. 14 November 2023; 2822 (1): 020020. https://doi.org/10.1063/5.017319

[26] Cascade Classification with Adaptive Feature Extraction for Arrhythmia Detection. J Med Syst (2017). 41: 11. DOI 10.1007/s10916-016-0660-9

[27] ECG Heartbeat Classification: A Deep Transferable Representation. https://arxiv.org/pdf/1805.00794.pdf

[28] De Lannoy, Gaël, et al. "Weighted conditional random fields for supervised interpatient heartbeat classification." IEEE Transactions on Biomedical Engineering 59.1 (2011): 241-247.

[29] De Lannoy, Gael, et al. "Weighted SVMs and feature relevance assessment in supervised heart beat classification." Biomedical Engineering Systems and Technologies: Third International Joint Conference, BIOSTEC 2010, Valencia, Spain, January 20-23, 2010, Revised Selected Papers 3. Springer Berlin Heidelberg, 2011.

[30] Mar, Tanis, et al. "Optimization of ECG classification by means of feature selection." IEEE transactions on Biomedical Engineering 58.8 (2011): 2168-2177.

[31] Park, Juyoung, Seunghan Lee, and Kyungtae Kang. "Arrhythmia detection using amplitude difference features based on random forest." 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015.

[32] Park, Juyoung, et al. "Cascade classification with adaptive feature extraction for arrhythmia detection." Journal of medical systems 41 (2017): 1-12.

[33] Mar, Tanis, et al. "Optimization of ECG classification by means of feature selection." IEEE transactions on Biomedical Engineering 58.8 (2011): 2168-2177.

[34] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[35] G. Doquire, G. de Lannoy, D. François, M. Verleysen, "Feature Selection for Interpatient Supervised Heart Beat Classification", Computational Intelligence and Neuroscience, vol. 2011, Article ID 643816, 9 pages, 2011. https://doi.org/10.1155/2011/643816

[36] Saenz-Cogollo, Jose Francisco, and Maurizio Agelli. “Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification.” Algorithms, vol. 13, no. 4, Mar. 2020, p. 75. Crossref, https://doi.org/10.3390/a13040075.