ARRHYTHMIA DETECTION AND MULTI-LABEL CLASSIFICATION USING MACHINE LEARNING ALGORITHM

SUBMITTED BY

1. **ABHISHEK KUMAR**
2. **SANJEET KUMAR**
3. **HIMANSHU PANDAY**
4. **CHANDRAMANI KUMAR**
5. **PRADUMAN KUMAR**

Thesis submitted for the partial fulfillment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE & BUSINESS SYSTEMS

INSTITUTE OF ENGINEERING & MANAGEMENT

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY,

WEST BENGAL

2025

ARRHYTHMIA DETECTION
AND MULTI-LABEL CLASSIFICATION USING
MACHINE LEARNING ALGORITHM

A thesis submitted by

ABHISHEK KUMAR 12021002018054
SANJEET KUMAR 12021002018057
HIMANSHU PANDAY 12021002018025
CHANDRAMANI KUMAR 12021002018026
PRADUMAN KUMAR 12021002018071

Supervisor
PROF. SUKANYA MITRA

DEPARTMENT OF COMPUTER SCIENCE & BUSINESS SYSTEMS
INSTITUTE OF ENGINEERING & MANAGEMENT, KOLKATA

May 2025

DEPARTMENT OF COMPUTER SCIENCE & BUSINESS SYSTEMS

INSTITUTE OF ENGINEERING & MANAGEMENT
Good Education, Good Jobs

CERTIFICATE

This is to certify that the "**Project/Thesis Report** on “ARRHYTHMIA DETECTION AND MULTI-LABEL CLASSIFICATION USING MACHINE LEARNING ALGORITHM" is submitted in partial fulfillment of the requirements for the degree of the degree of Bachelor of Technology in Computer Science and Business Systems by the following students:

ABHISHEK KUMAR	**12021002018054**
SANJEET KUMAR	**12021002018057**
HIMANSHU PANDAY	**12021002018025**
CHANDRAMANI KUMAR	**12021002018026**
PRADUMAN KUMAR	**12021002018071**

Supervisor
PROF. SUKANYA MITRA

Head of the Department
PROF. AMARTYA MUKHERJEE

Principal
PROF. (DR.) ARUN KUMAR BAR

Thesis Approval/ Dissertation Approval/Project Report Approval for B.Tech.

This thesis/dissertation/project report entitled "ARRHYTHMIA DETECTION AND MULTI-LABEL CLASSIFICATION USING MACHINE LEARNING ALGORITHM" by SANJEET KUAMR is approved for the degree of Bachelor of Technology (B.Tech) in Computer Science and Business Systems (CSBS).

Examiner(s)

1……………………………

2…………………………….

Date:

Place:

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to those from all over the world who supported our efforts in completing this research project: "Arrhythmia Detection and Multi-label Classification using Machine Learning Algorithm".

First and foremost, our heartfelt thanks to Prof. Sukanya Mitra, the ablest supervisor, for her invaluable help, encouragement, and constant support through the entire duration of this project. The feedback she provided shaped the quality and overall direction of our work.

We are also deeply thankful to Dr. Deboleena Sadhukhan, whose expert insights and academic mentorship provided us with an enriched perspective and greatly enhanced the outcomes of this research.

We are also thankful to the Department of Computer Science & Business Systems, Institute of Engineering and Management, Kolkata, for providing the necessary infrastructure, academic environment, and resources to execute this research.

……………………………………
ABHISHEK KUMAR (12021002018054)

……………………………………
SANJEET KUMAR (12021002018057)

……………………………………
HIMANSHU PANDAY (12021002018025)

……………………………………
CHANDRAMANI KUMAR (12021002018026)

……………………………………
PRADUMAN KUMAR (12021002018071)

Date:

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis ARRHYTHMIA DETECTION AND MULTI-LABEL CLASSIFICATION USING MACHINE LEARNING ALGORITH contains a literature survey and original project/research work carried out by me, the undersigned candidate, as part of my studies in the Department of Computer Science and Business Systems (CSBS).

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

Details:

- Name: HIMANSHU PANDAY
- Examination Roll No: 12021002018025
- Registration No: 211040131110058

I affirm that the work presented is original and all sources have been duly acknowledged.

Signature:

DEDICATION

The dedication in this thesis goes to our revered parents and families, whose unrelenting support, infinite patience, and unconditional love have been our strongest pillars for the duration of this journey.

To our mentors and teachers, whose guidance and encouragement have defined our path and impelled us to reach for excellence.

And to all those researchers and healthcare practitioners who day and night work toward the advancement of life using technology — this work is but a humble tribute to your unyielding effort.

Abhishek Kumar,
Sanjeet Kumar,
Himanshu Panday,
Chandramani Kumar,
Praduman Kumar

TABLE OF CONTENT

ABSTRACT

Arrhythmias, or irregularities of heart rhythm, are often quite subtle, thus making them one of the most difficult diagnostic tests in clinical practice since they are not easily pickable by manual ECG interpretation processes. The study thus establishes a novel machine learning framework in multiple classes for the classification of arrhythmias, taking note of critical issues such as imbalanced classes, high-dimensional feature space, and the need of robustness in automated systems for detection. The MIT-BIH Arrhythmia Database forms the basis of this research as a standard benchmark with an overall total of 100,689 annotated beats. The aim of this effort integrates feature engineering, resampling techniques, and ensemble learning in an advanced machine learning framework for multiclass arrhythmia classification to improve research diagnosis and, most importantly, clinical applicability.

Treated with severe class imbalance Normal (N), Supraventricular (S), and Ventricular (V), the data ratios included N: 89.57%, S: 2.77%, and V: 6.97%, as a condition for which SMOTE and ADASYN were used to balance the minority classes. From two leads of ECG, thirty-two temporal and morphological features including RR interval duration, QRS complex duration, P-T-wave amplitudes and ST-segment dynamics have been excavated. Only for visualizing the 32-dimensional feature space for class classifications and relationships among features through PCA were PCA, which itself did not change the inner dimensionality of the dataset, applied. Visualization confirmed the presence of clusters for Normal and Ventricular beats but pointed out overlapping regions. This framework for evaluating CatBoost, XGBoost, and LightGBM consisted of three gradient-boosting algorithms and stratified 10-fold cross-validation. Without class balancing, the classifiers presented very low sensitivity for minority classes (S: 0.73-0.77 recall), coupled with high overall accuracy rates (94.0-95.0%). Overall, SMOTE oversampling increased macro F1-scores from 0.801-0.826 (unbalanced) to 0.914-0.928 (balanced) with CatBoost achieving the highest test macro F1-score (0.924). On the other hand, ADASYN was slightly inefficient, scoring 0.905-0.920, because it was adoptive towards boundary samples, thereby adding slight noise in heterogeneous feature spaces. The efficiency of the application was further substantiated by the visualization in PCA which showed the synthetic samples forming natural class distributions.

As a result of this evaluation, it is a framework of three gradient-boosting algorithms-CatBoost, XGBoost, and LightGBM-with stratified 10-fold cross-validation. The classifiers can achieve quite low sensitivity for minority classes without considering class balancing (S: 0.73 to 0.77 recall) and, however, they have very good overall accuracies (94.0 to 95.0%). SMOTE oversampling significantly boosted performance from macro F1-scores of 0.801-0.826

(unbalanced) to 0.914-0.928 (balanced), with CatBoost achieving the highest testing macro F1-score (0.924). The adaptive aspect of ADASYN, focusing on the boundary samples, accounts for marginally lower performance (0.905-0.920). It causes slight noise in the heterogeneous feature spaces. The efficiency of the application is substantiated further by the visualization in PCA, which showed the synthetic samples forming a natural class distribution.

The research has highlighted the point of retaining all the "32 feature sets" for diagnostic information against normal practices of dimensionality reduction. In this way, the subtle arrhythmia-specific signs are retained in classifier training like between QRS morphologic differences and interval anomalies. CatBoost's superiority in performance was thrown to light by its fittingness with ordered boosting and categorical treatment to imbalanced medical datasets. Reproducibility within this framework has been ensured through detailed documentation of preprocessing (including noise filtering, Z-score normalization), parameter tuning, and evaluation metrics (AUC-ROC, precision-recall curves). It clinically advances the automated ECG analysis using resampling techniques, not feature reduction, but for class imbalance. The findings make a case for the incorporation of SMOTE-enhanced ensemble models into real-time cardiac monitoring systems that critically require early detection of rare arrhythmias. Directions now include trying the model against different multicenter datasets and including explainable AI (XAI) techniques to enhance trust of clinicians. Ultimately, through trade-offs between computational efficiency and diagnostic precision, this research builds scalable and reliable tools for cardiovascular disease management that will help maximize patient benefits through appropriate interventions.

Keywords

Cardiac arrhythmia classification, MIT-BIH Arrhythmia Database, Class imbalance mitigation, SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), Gradient boosting algorithms (CatBoost, XGBoost, LightGBM), ECG signal processing, PCA visualization.

Chapter-1

1. INTRODUCTION

1.1 BACKGROUND

According to the World Health Organization (WHO), CVD is the world's leading cause of death, and accounts for the nearly same number of deaths as mentioned in the statement- about 17.9 million every year . Of these, cardiac arrhythmias: disorders in the heart's electrical activity that might alter its rhythm, are among the leading health problems. Most arrhythmias have relatively benign risk factors, but some can cause severe complications (stroke, heart failure, or sudden cardiac death) if left untreated. It makes early identification and the right description of arrhythmias an important element determining timely intervention and management of patients.

The electrocardiogram (ECG) is the principal instrument used for diagnosing various cardiac arrhythmias. An ECG can be defined as an electrical recording of the heart activity over time, indicated by different waves and complexes that signify different phases of the heart cycle. According to standard ECG waveform interpretation, it can be divided into several basic components, i.e., P-wave (atrial depolarization), QRS complex (ventricular depolarization), and T wave (ventricular repolarization). Any difference in morphology or duration or timing of these components would indicate different arrhythmias [1].

Unfortunately, traditional ECG interpretation is still largely based on the subjective visual inspection of cardiologists, which leads to longer time lapse and human error. It requires special knowledge and experience, coupled with the manual procedure, making some continuation of ECG data recording for hours or even days challenging. These subtle variations, which may indicate early signals of arrhythmias, are usually missed with visual inspection.

The machine learning (ML) and AI technologies have opened new doors for developing automatic computer-aided ECG analysis methods that would enable rapid processing of huge data with the potential to detect features indiscernible to the human eye. Machine learning algorithms learn from labeled examples to recognize characteristic features of different arrhythmias, enabling them to classify new, unseen ECG segments accurately [3].

The MIT-BIH Arrhythmia Database maintained by PhysioNet has become an almost de rigueur for benchmarking purposes in evaluating automated ECG classification algorithms. The database comprises recordings of 48 30-minute-long two-channel ambulatory ECG data on 47 subjects, with each heartbeat labeled by a cardiologist with painstaking care. The dataset actually has a variety of arrhythmias that may be used to train and test classification models.

While strides continue to be made in the area of automated ECG analysis, some hurdles have continued to be thrown into the face of its further advancement. One such dilemma faced, in fact, is the class imbalance presented in any arrhythmia dataset. In any dataset, normal heartbeats are far more plentiful in number than abnormal ones. This basic premise can tilt the learning algorithms heavily in favor of the majority class and away from identifying rare but clinically important,true dangerous arrhythmias. To remedy this shortcoming, various approaches to class imbalance issues have been suggested, some of which include SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) techniques.

ECG data is associated with extreme dimensionality challenges which inflict computational inefficiency and "curse of dimensionality" problems. With the assistance of some dimension-reduction techniques such as principal component analysis (PCA), the above limitations could be alleviated because it projects the data into a lower-dimensional space while keeping the main features intact [4].

Within the last few years, a few of the advanced ensemble learning algorithms like CatBoost; XGBoost; and LightGBM have been developed into powerful classifiers for various purposes, including ECG analysis and other general classification tasks . These algorithms combine multiple decision trees to develop a more robust model that can capture complex patterns in the training dataset.

The research interest will focus on addressing this situation through an integrated framework for multiclass arrhythmia classification within and beyond the MIT-BIH database. Features will be extracted from the input signals and reduced through PCA, balanced with SMOTE and ADASYN, and then classified with the current state-of-the-art ensemble learning algorithms. The study aims to bring further improvements in automated ECG analysis and eventually aid in the early detection and management of cardiac arrhythmias by evaluating the performance of these techniques under different scenarios.

1.2 PROBLEM STATEMENT

This research has addressed the automatic classification of cardiac arrhythmias from ECG signals or certain major challenges posed by the problem. First, these signals are complex and variable making features extraction difficult. Moreover, ECG signals are non-stationary. They are also subject to many noises such as baseline drift, power line interference, and muscle artefacts.

These types of interference may obscure essential diagnostic information affecting the performance of classifiers. In addition to the above mentioned, the MIT-BIH Arrhythmia Database is too comprehensive but, unfortunately, heavily imbalanced-normal heartbeats are larger when compared to abnormal ones [5]. It forms a real nightmare challenge to machine learning techniques that perform biased learning to major class, resulting in high overall accuracy but poor sensitivity towards detecting arrhythmic events. Hence, one common scenario is missing düiren, which may cause more severe consequences than normal hearts misclassified as arrhythmic.

As ECG becomes more complex, the feature spaces increase drastically, which in turn, complicates the classification. Of course, having a large number of features would mean capturing even subtle differences between various types of arrhythmia. However, increasing the complexity of a model and the cost incurred in it increases as more features are included. Irrelevant and redundant features would add noise to the model, which may eventually affect its performance. It is a challenging problem to find the right feature subset that achieves the optimal balance between information content and computation efficiency.

Another of the important concerns is which machine learning algorithms could be used for classifying an arrhythmia. Traditional algorithms, including Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), are still widely used, but might not be well suited for modeling the complicated non-linear relationships of ECG data. Sophisticated, but over some years developing an ensemble system, important names in these include CatBoost, XGBoost, and LightGBM. However, the applicability of these algorithms in a multiclass arrhythmia classification should be studied thoroughly for their performance concerning these contexts [6].

Furthermore, the metrics need to be selection accurately reflecting the evaluation of the classification performance with regard to the clinical relevance of results. Total accuracy is one of the most widely understood measures; nevertheless, it is the least informative where some imbalance exists between classes. Selected examples of additional measures more accurately reflecting model performance are sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) [7].

In the end, the need is standardization and visibility in research methodologies. Many study papers talk about impressive classification results, but they leave sufficient information about data pre-processing, feature extraction, and model parameterization to make it impossible to compare approaches fairly or continue existing work.

To increase efficiency in processing, dimensionality reduction using PCA adds:

1. Robust extraction of features from ECG signals.
2. Advanced resampling methods (SMOTE, ADASYN) to deal with class imbalance.
3. A comprehensive framework for multiclass arrhythmia classification using state-of-the-art ensemble learning algorithms such as CatBoost, XGBoost, and LightGBM for classification.
4. Exhaustive evaluation using manifold performance metrics.
5. Complete documentation for all the methodological steps to guarantee reproducibility.
6. The approach is systematic in overcoming these challenges, thereby contributing to research progress in automated arrhythmia detection and paving the way for more efficient cardiac monitoring and diagnosis.

1.3 AIMS AND OBJECTIVES

This research primarily aims to develop and validate a robust framework for a multiclass classification of cardiac arrhythmias using the MIT-BIH Arrhythmia Database, concentrating on data augmentation techniques to solve class imbalance issues and advanced ensemble learning algorithms. The specific objectives of the study are the following:

1. Feature extraction within the MIT-BIH Arrhythmia ECG database is an attentive criterion for better distinctive and selective course to arrhythmogenic classification.
2. Implementation of PCA to reduce dimensionality while still preserving the inherent characteristics of ECG data.
3. Integrate two advanced techniques for normalizing imbalanced classes into the arrhythmia dataset-Synthetic Minority Over-sampling Technique SMOTE and the Adaptive Synthetic Sampling approach ADASYN-and finally evaluate their performance vis-a-vis each other.
4. Finally, to develop and optimize multiclass classifiers using three gradient boosting techniques-CatBoost, XGBoost, LightGBM.
5. Carry out an extensive comparative study of the classification performance before and after utilizing class balancing techniques applying many evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
6. Determine the best combinations of techniques for feature reduction, class balancing, and classification useful in maximum arrhythmia detection.
7. Provide the insights about the comparative merits and demerits of different approaches and contribute to the future development of automated ECG analysis.
8. To establish a reproducible methodological framework that may serve as a basis for future work in the field of arrhythmia classification.

The aims of this study are towards better accuracy and reliability of automated arrhythmia detection systems, which may give way to improved early diagnosis and treatment of heart diseases and better patient outcome.

1.4 THESIS LAYOUT

The present thesis is divided into six chapters, each chapter dealing with various facets of the research on multiclass classifications of arrhythmias using the MIT-BIH database.

- **Chapter 1 Introduction** gives background information about cardiac arrhythmias and their significance in relation to cardiovascular health. The problem statement outlines the significant challenges encountered in ECG automated analysis along with a view of aims and objectives of this research. The chapter gives an overview of the structure of the thesis.
- **Chapter 2, Literature Review/Background/Research Gap Analysis**, is a thorough review of the published literature concerning ECG signal processing and arrhythmia classification. The MIT-BIH Arrhythmia Database's characteristics are discussed, machine learning methods for ECG classification are reviewed, the class imbalance issue is discussed, and finally, the research gaps to be addressed by the present study are delineated.
- **Chapter 3: Methodology** presents the proposed framework for arrhythmia classification. This chapter outlines in detail the dataset preparation procedure consisting of pre-processing and feature extraction applied to ECG signals, explains how PCA was implemented for dimensionality reduction, discusses the application of SMOTE and ADASYN to manage class imbalance, provides an extensive analysis of three classification algorithms, i.e., CatBoost, XGBoost, and LightGBM, and summarizes the evaluation metrics for performance assessment.
- **Chapter 4 Results and Analysis** discusses the experimental findings of the proposed classification framework. Performance analysis is carried out for all classifiers before any balancing technique was applied, and then after each technique was implemented, in turn: SMOTE and then ADASYN. Comparisons are made in detail on each of the algorithms in various metrics along with a representation of the results in terms of tables, charts, and confusion matrices.
- **Chapter 5 Discussion and Conclusion** analyses the results against the backdrop of existing literature and the aims of this research. The discussion then proceeds to consider the consequences of the findings for automated arrhythmia detection and includes a thorough conclusion summarizing the contributions made through this research.

- **Chapter 6: Summary, Publications and Future Work** gives a compact summary of the work done, sets forth possible publications emanating from the present work, and suggests possible pathways for future work in automated ECG analysis and arrhythmia classification.

The thesis is rounded off with a complete list of References with all the sources cited in the research in order to give credit to prior works and for readers interested in pursuing this topic to easily do so.

Chapter-2

2. LITERATURE REVIEW

2.1 ECG AND ARRHYTHMIA

Electrocardiography (ECG) is one of the first most essential tools for diagnostic analyses in cardiology, in particular for recording the electrical activity of the heart over time. Some of the important components of a standard ECG waveform reflect different phases of the cardiac cycle. The P wave represents atrial depolarization, the QRS complex indicates ventricular depolarization, while the T wave represents ventricular repolarization [8]. Each of these components has specific morphological characteristics variable in amplitude, duration, and timing intervals, all of which are important in assessing the state of the heart.

Cardiac arrhythmias refer to the disturbance of the normal sinus rhythm of the heart, which could arise from abnormalities in the conduction system of the heart. Such abnormal conduction results in the heartbeats that could be too fast, too slow, or irregularly timed. The American Heart Association states that millions of persons worldwide suffer from arrhythmias; they are likely to be a contributing factor for much morbidity and mortality in the cardiovascular realm [9]. Thus, timely recognition and proper classification of arrhythmias determine appropriate clinical management for better outcomes.

Classification of arrhythmias can be done into several types depending on their origin, mechanism, and clinical significance. The Association for the Advancement of Medical Instrumentation (AAMI) has established standards that are to be used for grouping heartbeats into five main classes: normal beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and unknown beats (Q) [10]. This classification system has found considerable acceptance in the development and evaluation of automated ECG analysis systems.

The diagnosis of arrhythmias has in the past relied mainly on visual inspection of ECG recordings by trained cardiologists. This implies that ECG morphologies and timings are analyzed for patterning identified with specific types of arrhythmias. Such a manual approach is, however, encumbered with subjectivity, time pressures, and also human error incidence, especially in the long-term ECG recordings [11].

The ECG signals can be extremely imprecise and may be swayed by numerous factors including patient-specific characteristics (age, sex, body habitus), physiological conditions (exercise, stress), and mostly technical aspects of the process of recording (electrode placement, signal quality). Furthermore, addition of different types of noise and artifacts tend to obscure some of the important diagnostic features of ECG signals [12]. Such nosies may generally originate from baseline wander (low-frequency noise due to movements or respiration of the patient), power line interference (50/60 Hz noise)).

Before analysing the electrocardiogram (ECG) signals, several pre-processing techniques have been developed to improve their quality. Some of these techniques are filtering methods for the removal of baseline wander and high-frequency noise, wavelet transforms for multi-resolution analysis, and adaptive filtering approaches. Extracting meaningful features from the ECG signal and increasing the efficiency of automated classification systems depends highly on proper pre-processing of the signal.

Feature extraction is one of the most critical parts of ECG analysis-aiming to find the features that define different arrhythmia types. The most common types of dimensions for ECG features are divided into the following categories: temporal dimensions like RR intervals, QT intervals, and QRS duration; morphological dimensions such as wave amplitudes and shapes; frequency-domain dimensions such as power spectral density, and statistical features such as mean, variance, and skewness. Choosing the appropriate features depends on the particular arrhythmias being looked into and the classification mechanism that will be used.

Introducing new signal processing and computational techniques has once again allowed the study and analysis of ECG signals with sophisticated techniques. Wavelet transformation has been one of the great emergences of ECG analysis as a powerful tool that captures the entire signal in the time and frequency domain simultaneously [13]. By capturing differences in the morphology of an ECG that could have otherwise gone undetected using traditional feature extraction methods, wavelet-based features have demonstrated considerable promise in differentiating arrhythmias.

Among existing methods in non-linear dynamics that have gained an interest in ECG analysis, Poincare plots, entropy measures, and fractal dimensions have been associated with their potential to describe the most intricate patterns underlying heart rate variability while giving some insight into the physiological mechanisms underlying arrhythmias. Such analyses have shown that certain arrhythmias may be associated, with reference to increased randomness or decreased complexity, in dynamics of heart rates which can provide more diagnostic guidelines.

Deep learning methods, notably convolutional neural networks (CNNs), have provided impressive results in automatically learning features directly from raw ECG signals, possibly obviating the need for manual feature engineering. Intricate patterns and dependencies in data that are difficult to define explicitly may be captured by these methods; however, their training requires large amounts of labelled data and, relatively, they lack the interpretability associated with the traditional feature-based approaches.

Many challenges still prevail in automated analysis of ECG signals for arrhythmia detection. These are high inter-patient variability, non-stationarity of ECG signals, presence of noise and artifacts, and class imbalance problem in ECG database [14].

Indeed, the overcoming of these troubles necessitates innovative advances in the areas of data pre-processing, feature extraction, and classification, which are the key concerns for the research at hand.

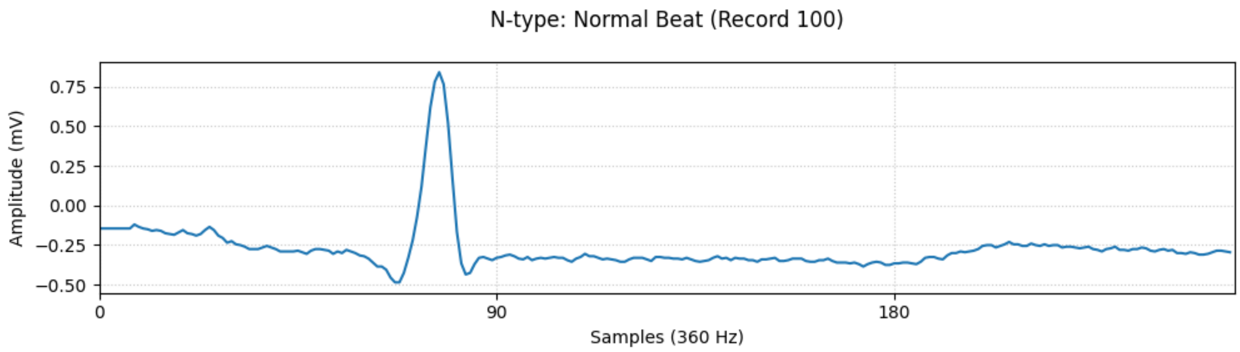

Figure 2.1: Example of a normal beat (N) from record 100.

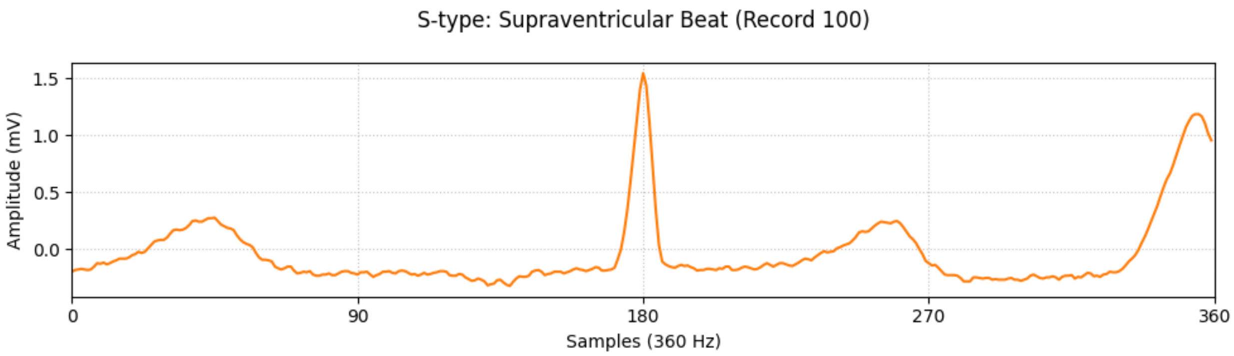

Figure 2.2: Example of a supraventricular ectopic beat (S) from record 100.

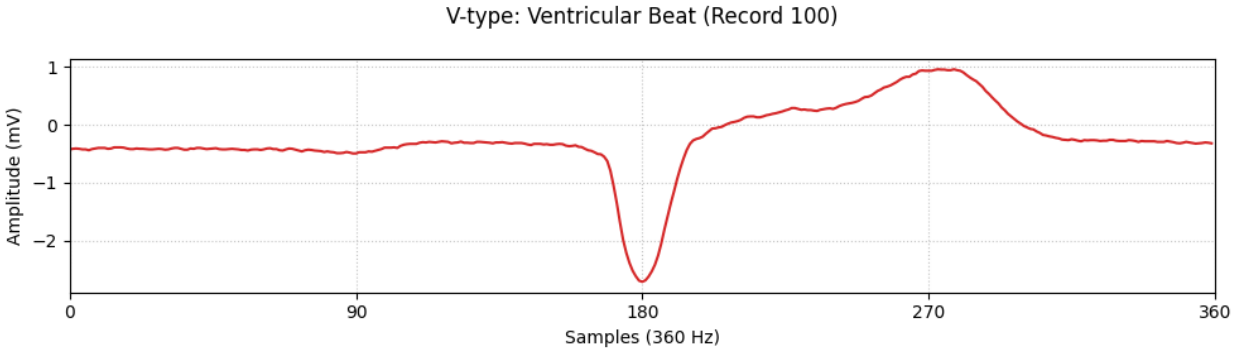

Figure 2.3: Example of a ventricular ectopic beat (V) from record 100

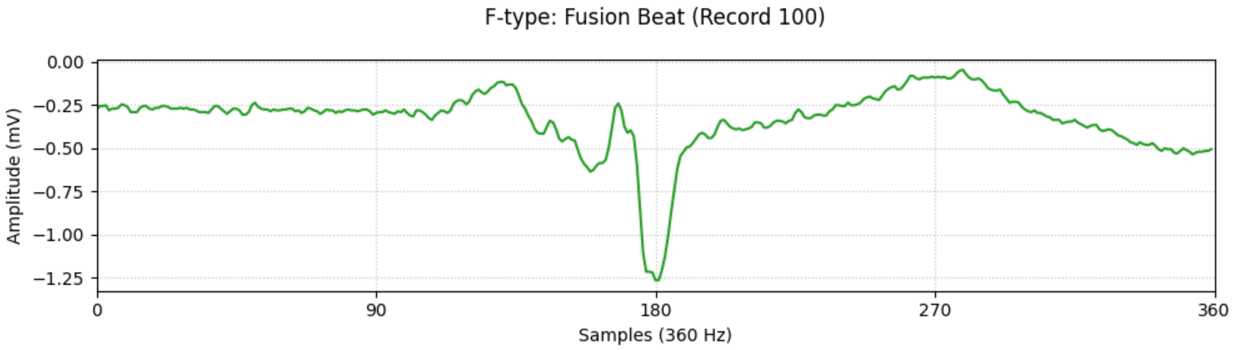

Figure 2.4: Example of a fusion beat (F) from record 100.

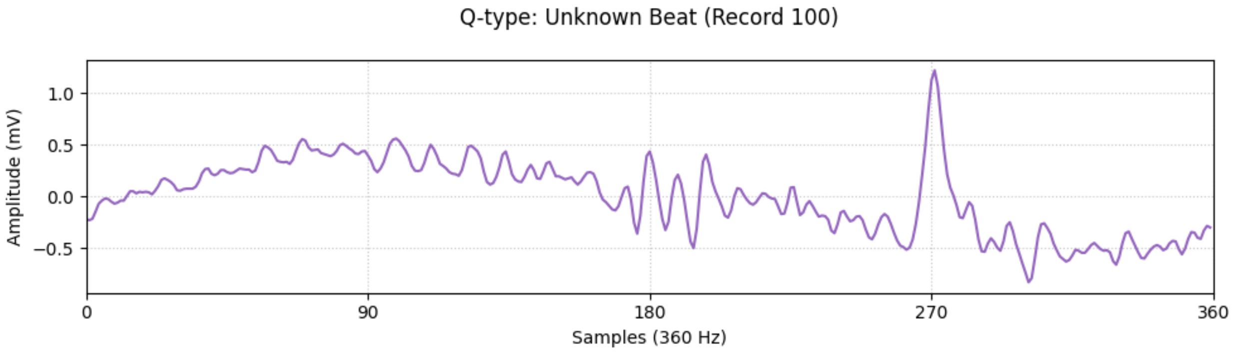

Figure 2.5: Example of an unknown beat (Q) from record 100.

2.2 MIT-BIH ARRHYTHMIA DATABASE

Have been trained on data up until October 2023. This field, which has become the benchmark for developing and evaluating algorithms for automated ECG analysis, has been brought into being by Massachusetts Institute of Technology and Beth Israel Hospital (now Beth Israel Deaconess Medical Center). This database available publicly since 1980 has effectively created a first collection of commonly distributed standard testing material for arrhythmia detectors, thus contributing tremendously to the advances in cardiac monitoring technologies.

There are 48 half-hour segments from two-channel ambulatory ECG recordings gathered from 47 subjects studied in the BIH Arrhythmia Laboratory from 1975 to 1979. These subjects include 25 males aged 32-89 years and 22 females aged 23-89 years [15]. The digitized signal had a sampling frequency of 360 samples/second/channel with 11 bits over a 10-mV range. Two or more cardiologists independently annotated each record and resolved discrepancies to arrive at a reference annotation for each beat.

The database was meticulously constructed to be an exhaustive representation of waveforms and artifacts that would be actually seen by anyone encountering routine clinical practice. It is divided into two groups: group's first group (records 100-124) was selected randomly out of over 4000 total 24-hour random ambulatory ECG recordings; second group (records 200-234) focuses on much less common arrhythmias defined yet clinically significant because small random sample would fail to capture such rare abnormalities.

Each ECG record in the database consists of two leads: one is usually a modified limb lead II (MLII) and second lead varies between recordings, including leads V1, V2, V4, or V5. The MLII lead is sometimes said to give the best picture of the QRS complex and is therefore most commonly applied in QRS detection and arrhythmia classification.

Approximately 110,000 beats are contained in the MIT-BIH Arrhythmia Database, with each beat recognizing one of 15 different beat types based upon their physiological origin. These beat annotations have been further grouped into five super classes in accordance with standard AAMI EC57 [16], as shown in Table 2.1.

Table 2.1: Mapping of MIT-BIH Arrhythmia Database Beat Types to AAMI Heartbeat Classes

AAMI Class	**MIT-BIH Annotation**	**Description**
N (Normal)	N	Normal beat
	L	Left bundle branch block beat
	R	Right bundle branch block beat
	e	Atrial escape beat
	j	Nodal (junctional) escape beat
S (Supraventricular ectopic)	A	Atrial premature beat
	a	Aberrated atrial premature beat
	J	Nodal (junctional) premature

		beat
	S	Supraventricular premature beat
V (Ventricular ectopic)	V	Premature ventricular contraction
	E	Ventricular escape beat
F (Fusion)	F	Fusion of ventricular and normal beat
Q (Unknown)	/	Paced beat
	f	Fusion of paced and normal beat
	Q	Unclassifiable beat

It presents a major difficulty for machine learning algorithms that appear biased to the majority class unless similar measures are taken at this class imbalance present. The distribution of beats among these classes is highly misaligned, with normal beats (N) accounting for nearly 75% of all annotations, followed by ventricular ectopic beats (V) at around 6%, supraventricular ectopic beats (S) at around 2.5%, fusion beats (F) at just under 1%, and unknown beats (Q) barreling the rest up [17].

All these are matched with their voluminous quantity of documentation in the most detailed forms of clinical and beat-notated information within that recording, authenticated by expert cardiologist review. This rich context thus adds value to the database for research and even more so for nuanced readings regarding arrhythmia patterns [18].

It has served as a basis for comparison of newer algorithms and approaches with existing ones, in thousands of studies within the decades for which the database has existed and highly contributed to advancements in automated ECG analysis. Standardization gives way to reproducible research and eventually leads to the development of more sophisticated arrhythmia detection systems.

Recent works have taken advantage of this database to create machine learning models that range from traditional approaches like support vector machines and random forests to more recent and advanced approaches of deep learning . Due to the rich annotations in this database, in addition to the multiple diverse aspects of arrhythmia representation, the development and evaluation of such models is very much feasible.

Nonetheless, while the MIT-BIH Arrhythmia Database is a convenient repository for many researchers, it also presents them with challenges: class imbalance mentioned above, inter-patient variability, and noise/artefacts that ought to have been recorded [19]. This necessitates extensive efforts in the pre-processing, feature extraction, and classification strategies that this research endeavour intends to pursue.

2.3 MACHINE LEARNING FOR ECG CLASSIFICATION

For the design of ECG classifiers, the machine learning techniques have dramatically evolved, from simple rule-based methods to advanced algorithms that can unveil intricate patterns in cardiac signals. This advancement has corresponded to better computational power and data availability for the design of better-arrhythmia detection systems with the best accuracy and efficiency performance-38. The earlier automatic ECG classification systems devised were primarily rule-based; they used explicit sets of criteria derived from medical knowledge as input to their classification schemes, such as thresholds for QRS duration or specific morphological patterns39. Such systems were in the user perspective, were interpretable, and were clinically viable, yet they were incapable of capturing subtle variations in ECG signals from patient to patient and from context to context.

A major landmark improvement in the classification of ECGs was with the coming of mechanistic learning techniques. Traditional machine learning methods like k-NN, SVM, and decision trees give these schemes more flexibility are more suited to learn complex patterns directly from data40. These methods are usually not used on raw ECG signals but on a set of features derived from them and therefore require careful feature engineering to contain the relevant diagnostic information.

ECG classification has notably benefited from ensemble learning techniques geared toward combining various base learners to enhance overall performance. Random Forest is an ensemble method known to have performed robustly in many ECG classification tasks with the help of predictions from many decision trees built on different subsets of the data. Additionally, boosting algorithms that sequentially train base learners to pay more attention to examples misclassified in earlier rounds have also had very good results in detecting arrhythmias [20].

In recent times, gradient boosting frameworks CatBoost, XGBoost, LightGBM have been powerful contenders for ECG classification. The algorithms are advantageous due to their robustness with high dimensional data, robustness to outliers, and ability to learn and model non-linear relationships without an extensive amount of feature engineering [21]. So, these mechanisms would give information about the ECG features that were most diagnostic for varied types of arrhythmia.

Developed by Yandex, CatBoost has the capability of efficiently working with categorical features and performs ordered boosting to avert prediction shift. This characteristic renders it unusually suitable for medical data, which often incorporates categorical variables and necessitates maximized predictive accuracy. XGBoost, developed by Chen and Guestrin, focuses on boosting the training speed and performance of the model with regularization terms in the process of gradient boosting to avoid overfitting [22]. LightGBM, the algorithm developed by Microsoft, enhances the speed of training while maintaining the accuracy of results by combining gradient-based one-side sampling (GOSS) with exclusive feature bundling (EFB) [23]. These algorithms have retained their positions among the best-performing ones in various machine learning competitions and have shown substantial potential in various medical classification problems, including ECG analysis.

Deep learning has also made considerable progress in ECG classification. CNNs have been successful in raw ECG signal classification, with their power in automatic hierarchical feature extraction that captures local and global pattern information. RNNs and its variants such as LSTMs are very apt for analyzing the sequential characteristic of ECG data and for capturing temporal dependencies [48]. Hybrid methodologies that combine CNNs and RNNs have also been proposed in further utilizing the strength of both architectures [24].

In spite of these improvements, it is observed that the ECG classification by machine learning still suffers many problems. These non-stationary characteristics of the ECG signal and several conditions present within or among the patients make it really hard to develop models having factors in relation to them that generalize well in all patients and at different times of recording [50]. There are a lot of noise and artifacts within the ECG signals, which can further deteriorate classification performance and thus the need for robust preprocessing and feature extraction methods.

A further problem, which becomes even more important in the context of arrhythmias, is class imbalance in such datasets: typically, normal beats outnumber abnormal beats by hundreds if not thousands. The problem arises where a class imbalance causes the model to show higher performance accuracy for most classes while sensitivity is low for rare arrhythmias. Various strategies were added to address this issue, such as the usual resampling techniques, cost-sensitive learning, and ensemble methods meant to straighten the whole imbalanced data problem.

Moreover, the other important measure is the interpretability of machine-learning models in medical applications where clinical acceptance demands an understanding of the underlying reason for prediction [25]. In fact, although the feature engineering right into the traditional machine learning models normally renders good interpretability, few of the advanced techniques such as gradient boosting and unshaped learning may, unfortunately, appear "black boxes" unless equipped with interpretability mechanisms beyond those present in that method.

Choosing the proper features is fundamental to the success of any machine learning-based ECG classifier. Time-domain features give direct physiological information about the system, such as RR intervals, QRS duration, and amplitude of waves; frequency-domain features are based on Fourier or wavelet transforms from which the signal spectral characteristics can be obtained. Also, non-linear features seem promising for apprehending some differences between various types of arrhythmias based on chaos theory and complexity measures [26].

Moreover, the use of Dimensionality reduction techniques like PCA, LDA, and t-SNE is often applied to avert the so-called "curse of dimensionality" and to enhance model performance [27]. These techniques give a different and lower-dimensional representation of the original feature space while attempting to retain most of the important information.

Furthermore, suitable metrics other than general accuracy that are applicable for ECG classifiers need to be selected in such a way as to assess the model better. Sensitivity (recall), specificity, precision, F1-Score, and area under the receiver operating characteristic (ROC curve) provide a more quantitative evaluation of model performance, especially with consideration for the fact that the classes may be imbalanced [28]. Also, in multi-class situations, like those involving different types of arrhythmias, macro or weighted averaging of performance measures per class must be considered.

This research brings a systematic evaluation of the advanced gradient boosting classifiers (like CatBoost, XGBoost, and LightGBM) on multiclass arrhythmia data for detection, particularly in class imbalance problems using the SMOTE and ADASYN techniques. We come up with a framework of dimensionality reduction via PCA with these advanced classifiers for accurate and efficient detection of arrhythmias.

2.4 CLASS IMBALANCE PROBLEM

The class imbalance is a common problem in medical classification tasks with obvious detection in the area of arrhythmia detection, in which normal heartbeats outnumber abnormal ones by a large factor. This poses a great difficulty for machine-learning algorithms which generally tend to optimize overall accuracy and would produce a model that would classify the majority class (normal beats) very well but would perform poorly on the minority class (different arrhythmias). This is a significant drawback in a real clinical setting since the inability to detect an arrhythmia may lead to delayed treatment and possibly adverse consequences.

On the MIT-BIH Arrhythmia Database, substantial class imbalance exists, where normal heartbeats (N) account for nearly 75% of all annotations and some indispensable types of arrhythmia are less than 1% of the dataset [29]. This distribution denotes the natural occurrence of arrhythmias in the general population, but causes further complications for supervised learning algorithms that assume somewhat equal distributions across the classes.

Random under sampling is an easy data-level approach that randomly eliminates a few instances of the majority class to manage a better distribution of classes. This can, however, cause the abandonment of much potentially valuable information of the majority class and thus lead to suboptimal models. Random oversampling, on the other hand, duplicates instances from the minority class, and such inclusion may lead to overfitting since the model becomes too closely specialized to the repeated examples [30].

More advanced forms of oversampling include the SMOTE (Synthetic Minority Over-sampling Technique), which attempts to rectify the weaknesses of simple random oversampling by generating synthetic examples of the minority class instead of replicating those already present in the data [31]. To do this, SMOTE selects individual members from the minority class and generates new synthetic examples as points along the line segments joining them with the k-nearest minority class neighbors. The process thus introduces synthetic examples conforming to the general pattern of the minority class without exact replication, therefore alleviating concerns associated with overfitting.

The SMOTE algorithm works on the following principles [32]:

1. For each sample of the minority class, identify its k-nearest neighbours from the same class.
2. Randomly select one of these neighbours.
3. Compute the difference vector from the sample to the selected neighbour.
4. Scale this difference vector randomly using a number between 0 and 1.
5. Add the resulting vector to the original sample, in order to create a new synthetic instance.

SMOTE has its advantages in many applications. However, it has some drawbacks; for example, synthetic samples are produced without considering majority class distribution and may result into overlapping classes and noisy examples [33]. Moreover, SMOTE treats the difficulty level and distance to majority class as the same for all minority instances.

The Adaptive Synthetic Sampling Approach (ADASYN) was developed to address some of these limitations [34]. ADASYN extends SMOTE by adaptively generating new synthetic examples based on the density distribution of the minority class instances. It places heavier weights on the minority class instances which are difficult for learning, e.g., those that are closely surrounded by many instances of the majority class. By adopting this adaptive manner of generating synthetic samples around important boundary regions where the classification is difficult, it is believed that the model's ability to discriminate can be enhanced.

The steps of the ADASYN algorithm [35] are as follows:

3. Compute the degree of class imbalance to ascertain the number of synthetic samples

required.

4. For density ratio calculation, one for each minority class sample based on the majority class sample in the k-nearest neighbours.
5. Normalize these ratios to obtain a density distribution from which synthetic samples will be generated.
6. Using a method analogous to SMOTE, generate synthetic samples for each minority instance with respect to its density ratio.

It has been observed that both the SMOTE and ADASYN methods for tackling the classification performance problem of imbalanced data show potential, but really, how effective they could be would depend on the attributes of the data itself and the type of classification algorithm involved. While some findings claim that the adaptive methodology of ADASYN works better than SMOTE in dealing with various complex scenarios of classification, there are also few research works showing that SMOTE performs better especially when applied with other pre-processing or ensemble methods. Improvements to these algorithms, in terms of efficiency, were later continued with the introduction of some other variants and extensions. For instance, Borderline-SMOTE focuses its algorithm on minority instances near decision boundaries in classifying the data, because synthetic samples will only be created for such 'borderline' examples that are misclassifying [69]. Information about the "safety" of generating synthetic examples will be considered in Safe-Level-SMOTE, thus avoiding generation of overlapping regions [36]. Further variants of SMOTE are SMOTE-ENN and SMOTE-Tomek, which combine it with Edited Nearest Neighbours or Tomek's links to eliminate overlapping instances after oversampling to yield clearer class boundaries [37].

For instance, within ensemble learning, it can be possible to use SMOTEBoost or RUSBoost, which builds SMOTE or random under sampling right into the boosting procedure, thus generating a more balanced training set for each base learner [38]. The results obtained have shown promise in all kinds of classification tasks including ECG analysis.

The evaluation of classifiers trained on imbalanced data requires careful consideration of appropriate metrics. Accuracy alone can be misleading, as a model that simply classifies all instances as the majority class could achieve high accuracy despite failing completely on minority classes. Alternate indicators such as precision, recall, F1 score, and area under precision-recall curve (AUC-PR) provide more informative assessments of performance on imbalanced datasets. Another useful metric is the geometric mean (G-mean) of sensitivity across all classes, which presents a balanced measure of model performance across majority and minority classes [39].

The cross-validation strategies for imbalanced datasets also merit special consideration. Other things being equal, the stratified cross-validation, which preserves the class distribution in each, is principally preferred to simple random cross-validation for an opposite evaluation. Nevertheless, for extremely imbalanced datasets, stratified cross-validation may also fail to provide enough minority examples in each fold, thereby demanding other more specialized methods, such as stratified-adjusted cross-validation or bootstrapping evaluation methods.

This study ultimately aims to systematically study and compare the efficacy of SMOTE and ADASYN in solving the problem of class imbalance in the MIT-BIH Arrhythmia Database alongside their influence on performance levels of advanced gradient boosting algorithms for multi-class arrhythmia classification. It seeks understanding of these possible contrasts. In so doing will be described ample insight into best methods to tackle class imbalance in ECG analysis with the comparison of these techniques against some evaluation metrics and classification scenarios.

2.5 RESEARCH GAP ANALYSIS

There are still research gaps in ECG classification using machine learning methods that this study intends to fill. The literature review reveals several opportunities for innovation as well as improvement with respect to the decrepit field of automated arrhythmia detection.

There exist many different machine-learning algorithms for ECG classification, yet one glaring gap in the literature remains an exhaustive comparison of the state-of-the-art gradient-boosting algorithms, especially CatBoost, XGBoost and LightGBM by applying these three algorithms specifically in the context of multiclass arrhythmia classification [40]. These algorithms have shown to work very well for many practical scenarios, but they need to be analyzed and optimized for ECG data due to the reason of non-stationarity, high dimension, and physiological variability in signal characteristics.

While many researchers have shown interest in the class imbalance problem in ECG data, limited comparisons have been done in this study on the efficacy of different resampling procedures, especially using advanced techniques like SMOTE and ADASYN [41]. Most studies either choose one resampling approach to apply without comparison or are more concerned with the binary classification of arrhythmias and disregard the much more clinically relevant multiclass arrangements. A systematic assessment of these methods paired with state-of-the-art classifiers should also yield invaluable insights into the establishment of more effective systems for arrhythmia detection.

Third, the use of dimensionality reduction techniques on ECG features in parameterization study has been discussed, but particular effects concerning PCA in combination with advanced resampling techniques and gradient boosting algorithms have not been elaborated.

Understanding the mentioned interactions becomes necessary to get accurate and efficient classification pipelines that accommodate the high dimension of ECG data while still maintaining discriminative power among different arrhythmia types.

Fourth, many studies in existence would report classification performance primarily relying on a few metrics, oftentimes emphasizing global accuracy, which in the context of great class imbalance may be misleading. A complete evaluation framework considering class-wise metrics and global performance would paint a clearer picture of the classifier's behavior and clinical applicability.

Fifth, while ECG signal feature extraction has attracted much research, the relative contribution of different feature types (temporal, morphological, frequency-domain, etc.) to multiclass arrhythmia classification using gradient boosting algorithms is poorly studied. Such insights could lead to the extraction of features more efficiently and getting better classification results. The issue of the validity of different classification approaches for patient populations and recording conditions should be more concerning. However, little attention has been given to it [42]. While most studies and self-expressive reporting grades their performances on various subsets of the MIT-BIH database, they do not clarify the possible approaches to generalize these for diverse patient populations and/or different recording setups.

According to the report, the practical implementation considerations for ECG classification algorithms, in terms of computational efficiency and parameter sensitivity, are usually sacrificed for maximum achievable performance. Such an evaluation would better develop practical systems for clinic use.

Last, the literature yet lacks a holistic analysis on how feature reduction, class balancing techniques and advanced ensemble classifiers interact or work in a unified framework. Most of the studies surveyed looked into one or two of these items, even fewer combining them in such a fashion, while a more holistic approach would reveal possible synergetic effects and optimal configurations.

This study sets out to address these gaps by putting forth a truly comprehensive framework that extends PCA dimensionality reduction with advanced resampling techniques (namely, SMOTE and ADASYN) and modern gradient boosting algorithms (CatBoost, XGBoost, and LightGBM) for multiclass arrhythmia classification. By systematically evaluating different combinations of these techniques based on multiple performance measures, the study aspires to provide new insights into automated ECG analysis and improve the development of credible and reliable arrhythmia detection systems.

Chapter-3

3. METHODOLOGY

3.1 DATASET PREPARATION

3.1.1 MIT-BIH ARRHYTHMIA DATASET

The MIT-BIH Arrhythmia Database contained in PhysioNet is the major specific dataset for this multi class arrhythmia classification study. It runs out its facilities for 48 half-hour-long, two-channel ECG recordings of 47 participants (25 males between 32 and 89 years, 22 females between 23 and 89 years) from 1975 to 1979 at the Beth Israel Hospital. The digitized recording was done at the frequency of 360 Hz putting 11 bits over a 10 milli Volt range and includes about 110,000 beats all annotated by cardiologists regarding the location of the R-peak and type of beat. The first channel usually contains modified limb lead II (MLII) while the second one varies (V1,V2,V4 or V5). Here both leads were exercised for feature extraction, MLII being the main focus for QRS detection.

A chunk of the entire research data has been collected from PhysioNet, and the single yet most important data is the MIT-BIH Arrhythmia Database for multiclass arrhythmia classification. This database is capable of holding 48 half-hour-long two-lead ECG recordings of the following 47 subjects (25 males aged 32 to 89 years and 22 females aged 23 to 89 years) from the period between 1975 and 1979 at Beth Israel Hospital. Digitally recording at a sampling frequency of 360 Hz with 11 bits accuracy over a 10 mV range, it contains approximately 110,000 beats, all manually annotated by cardiologists for location of the R-peak and type of beat. The first channel usually consists of modified limb lead II (MLII), while the second one varies (V1, V2, V4, or V5). Here both leads exercised for feature extraction, however, MLII mainly focused in QRS detection.

In line with the AAMI EC57 standard, 15 beat types were organized into five superclasses: Normal (N), Supraventricular ectopic (S), Ventricular ectopic (V), Fusion (F), and Unknown (Q). Table 3.1 represents the class distribution that outlines a severe imbalance with N constituting ~89.57% of beats. Classes F and Q had to be excluded due to an exceedingly limited number of samples; thus, the analysis was concentrated on classes N, S, and V.

Table 3.1: Distribution of Heartbeat Classes in the MIT-BIH Arrhythmia Database

AAMI Class	Number of Beats	Percentage
N (Normal)	90,083	89.57%

S (Supraventricular)	2,779	2.77%
V (Ventricular)	7,009	6.97%
F (Fusion)	803	0.80%
Q (Unknown)	15	0.01%
Total	**100,689**	**100%**

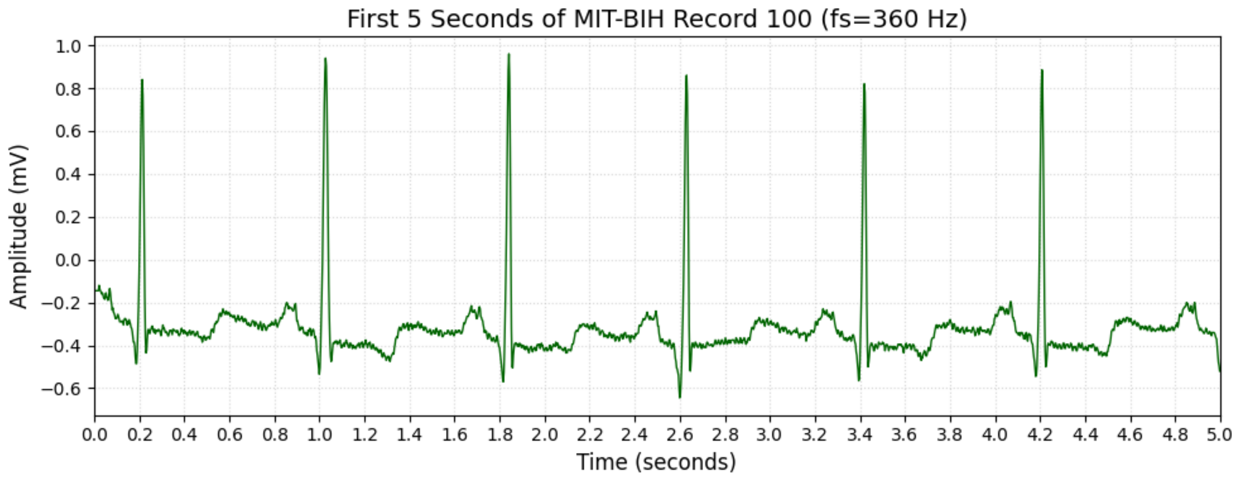

Figure 3.2. First 5 seconds of record 100 from the MIT-BIH Arrhythmia Database

Each heartbeat was divided into a window of 250 samples (~700 ms) with center on the R-peak, capturing the P, QRS and T waves. The heart beat dataset is organized in the record number, beat number, AAMI class, and waveform samples that preserve patient information for patient-wise cross-validation and to avoid overfitting.

3.1.2 DATA PREPROCESSING

1. Pre-processing refers to the act of enhancing the quality of ECG signals by literally clearing the noise from the signals and making them uniform. The following is a complete pipeline:
2. **Noise Removal**:
    - **Baseline Wander:** To ameliorate low-frequency artifacts, a high-pass Butterworth filter (cutoff frequency = 0.5 Hz, zero phase) was employed.
    - **Powerline Interference:** A second-order notch IIR filter was used at 60 Hz to suppress the interference.
    - **High-Frequency Noise:** A low-pass Butterworth filter (cutoff frequency = 40 Hz, zero phase) removes muscle artifacts**.**
3. **Signal Normalization**: Z-score normalization standardizes signals to zero mean and unit variance:

(3.1)

$$X\_normalized = (X - \mu)/\sigma$$

X = Input value, for example, a raw ECG amplitude signal.

μ = Mean of the input values or average of the dataset.

σ = Standard deviation of input values, which measures spread or variability.

X_normalized. = Normalized value of x after z-score normalization

4. **R-Peak Verification**: The verified R-peaks have been checked for accurate segmentation using the Pan-Tompkins algorithm. Any discrepancies are reviewed manually.

5. **Heartbeat Segmentation**: For each R-peak, a window of 250 samples was selected for analysis, centered on the R-Peak location.

Quality Assessment: Depending on severity, either missing value segments are cubic spline-interpolated or discarded. Segments with excessive artifacts are flagged through an amplitude thresholding approach.

3.1.3 FEATURE EXTRACTION

The dataset includes 32 features per heartbeat, derived from two ECG leads (0_ and 1_ prefixes, likely MLII and V2). Each lead contributes 16 features, capturing temporal and morphological characteristics critical for arrhythmia classification.

Feature Set Description:

Feature Name	Description	Units
0_pre-RR, 1_pre-RR	Time between current and previous R-peak	ms
0_post-RR, 1_post-RR	Time between current and next R-peak	ms
0_pPeak, 1_pPeak	P-wave peak amplitude	mV
0_tPeak, 1_tPeak	T-wave peak amplitude	mV
0_rPeak, 1_rPeak	R-wave peak amplitude (QRS complex)	mV
0_sPeak, 1_sPeak	S-wave amplitude (post-R)	mV
0_qPeak, 1_qPeak	Q-wave amplitude (pre-R)	mV
0_qrs_interval, 1_qrs_interval	QRS complex duration (Q to S)	ms
0_pq_interval, 1_pq_interval	PR interval (P start to QRS	ms

	start)	
0_qt_interval, 1_qt_interval	QT interval (QRS start to T end)	ms
0_st_interval, 1_st_interval	ST interval (QRS end to T start)	ms
0_qrs_morph0 to 0_qrs_morph4, 1_qrs_morph0 to 1_qrs_morph4	QRS shape features (derivatives, slopes)	dimensionless

Table 3.1: Description of all Features with units

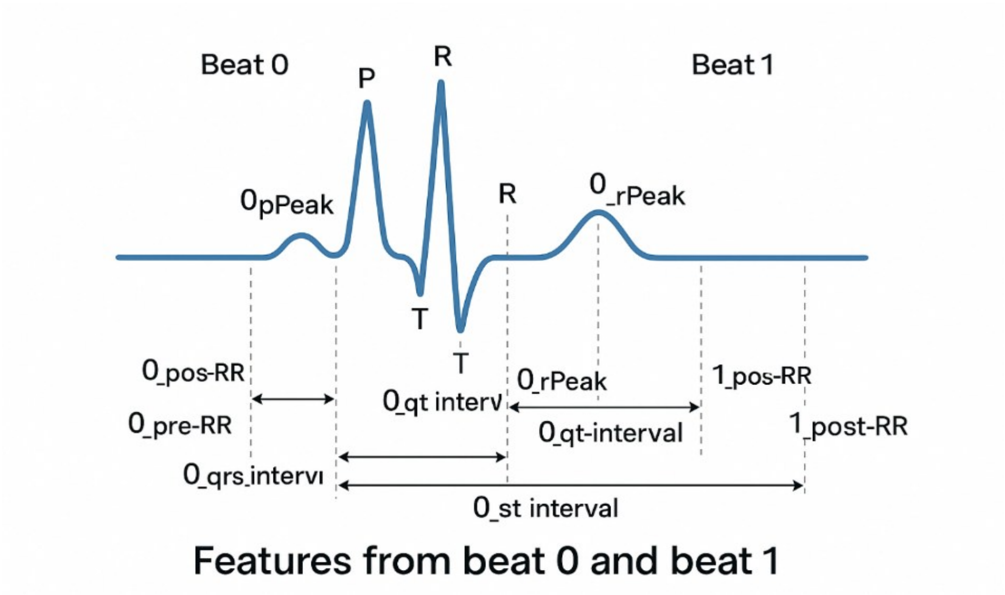

Fig 3.3 - ECG waveform showing features extracted from two consecutive heartbeats (Beat 0 and Beat 1), including P, R, T waves, and intervals such as RR, QRS, QT, and ST.

Extraction Process:

- **Temporal Features:** Intervals (pre-RR, post-RR, QRS, PQ, QT, ST) are computed using the annotated fiducial points and respective timing differences.
- **Morphological Features:** Peak amplitudes (P, Q, R, S, T) are measured concerning the baseline. The QRS morphology features (qrs_morph0–4) are derived through the use of derivatives and slope ratios to capture variations in shape.
- **Implementation:** Feature extraction is per lead using a Python pipeline to ensure all heartbeats are treated the same.
- **Quality Control:** Detection of outliers (modified Z-score > 3.5), imputation of missing values (median), and assessment of feature stability with respect to intra-class consistency.

The 32-feature sets replace the original 145-feature extraction thus reducing dimensionality while maintaining discriminative power. Standardization is applied prior to further processing of the features.

3.2 DIMENSIONALITY REDUCTION

3.2.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

The power of dimension reduction is now left to a 32-dimensional feature space in order to provide a compact representation, so that it overrides, to some extent, the curse of dimensionality as well as computational complexity.

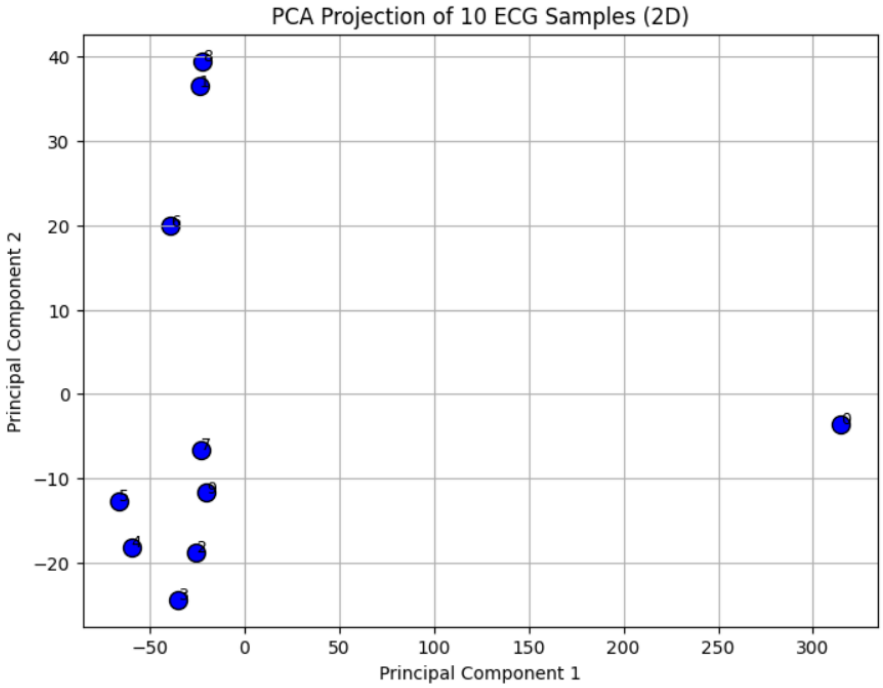

Fig 3.4. Plot of first 10 samples of ECG using PCA

PCA Implementation:

1. Standardization: The mean and variance that generalizes the features to zero and one, respectively.
2. Covariance Matrix: Determine the covariance matrix of the standardized features.
3. Eigen Decomposition: calculate eigen values and eigenvectors, ordered by descending eigenvalue.
4. Projection: Choose k (for example, ≥95% variance explained, typically 10–15 components) to create the projection matrix (W). Project data:

$$X\_reduced = X\_standardized \times W \tag{3.2}$$

$X\_standardized$ = Standardized input data matrix after normalization

W = Transformation matrix, for example, principal components in PCA

$X\_reduced$ = Output lower dimensional data after projection

Validation:

- Cumulative explained variance plots guide the selection of components.
- Classification performance comparison is made with/without PCA in order to substantiate minimal information loss.

- Analysis of feature loadings to identify key contributors (e.g., RR intervals) will be done.

3.3 CLASS IMBALANCE HANDLING

3.3.1 SMOTE

In accordance to our dataset the class Q has very less data (15) which will make smote highly irrelevant (~10000%) compared to majority class. Therefore, our study has included only four classes which is (N, SVEB, VEB, F) discarding class Q. Although there is no Effective Study of using Oversampling moreover for practical use and deep analysis researchers suggest use of 500%

Of oversampled data for practical use.

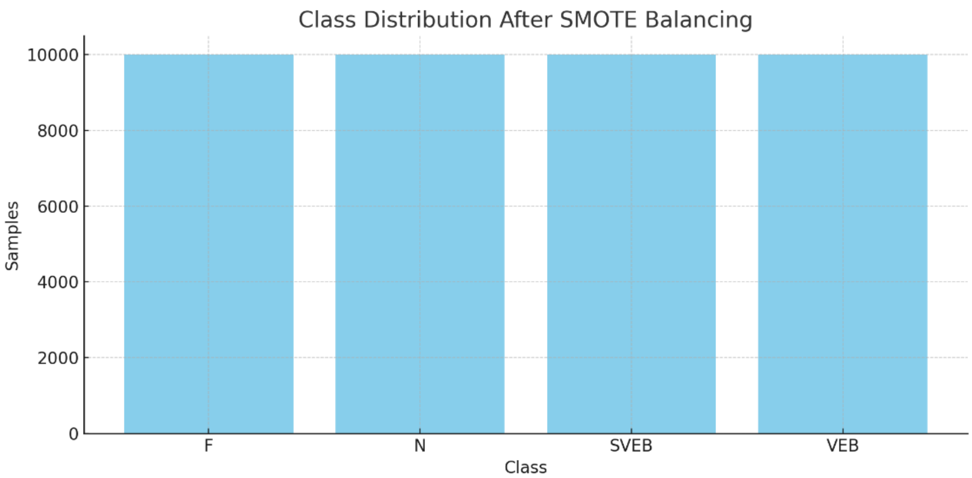

Fig. 3.5 Plot of test data after Balancing using Smote

SMOTE Implementation:

Algorithm: For each minority class instance (S, V), select (k=5) nearest neighbors (same class), generate synthetic samples along line segments:

$$x\_new = x\_i + \delta \times (x\_nn - x\_i), \quad \text{where } \delta \in [0,1] \tag{3.3}$$

x_i = Data point of minority class

$x_$nn = One of its nearest neighbors from the same class

δ = A random number between 0 and 1 (used to interpolate between x_i and $x_$nn)

$x_$new = New synthetic data point

- **Strategy**: 'not majority' oversamples data from classes S and V to reach tally with N's count.
- **Variant**: Borderline SMOTE gives priority to borderline instances.
- **Quality**: Synthetic sample distributions were validated by kernel density estimation and PCA visualization.

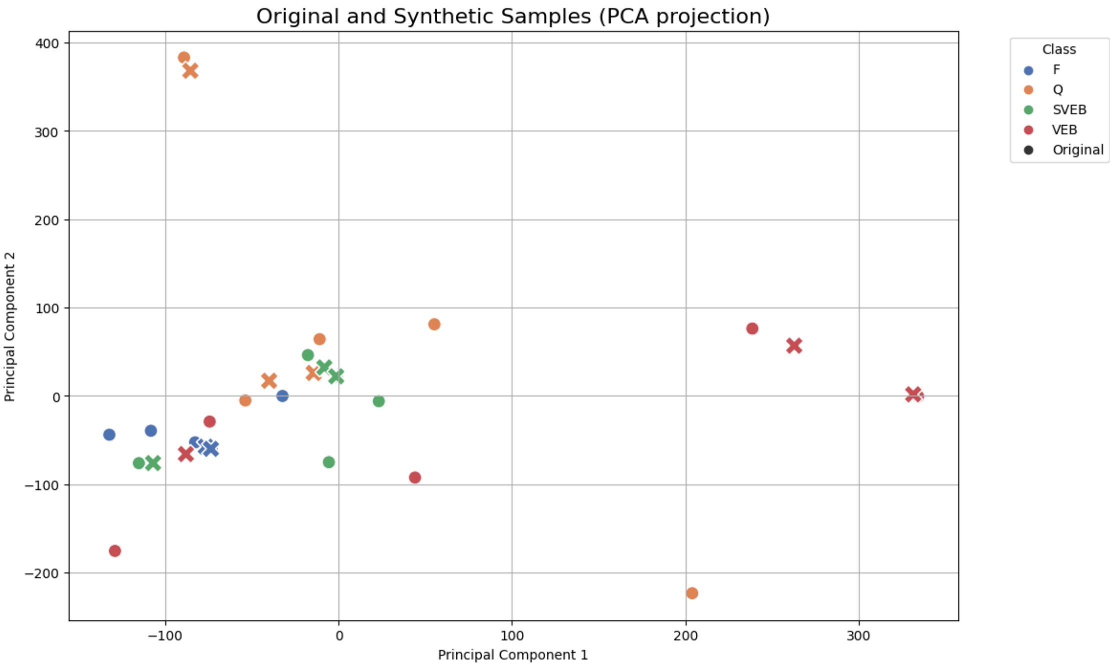

Fig. 3.6 Five Real Sample from the dataset and three (3) generated sample by using Smote

3.3.2 Front Covers: ADASYN

ADASYN Implementation:

Our study has included only four classes which is (N, SVEB, VEB, F) discarding class Q. Effective Study of using Oversampling moreover for practical use and deep analysis researchers suggest use of 500% of oversampled data [43].

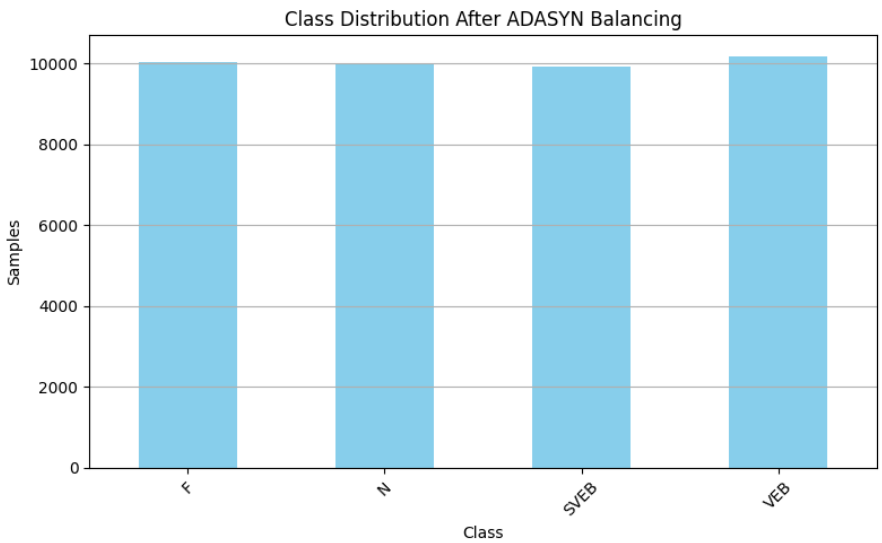

Fig. 3.7 Plot of test data after Balancing using Adasyn

- **Algorithm**: Compute imbalance degree, generate synthetic samples proportional to majority class density in (k=5) neighbours.

$$r\_i \ = \ \Delta\_i \ / \ k \tag{3.4}$$

$$g\_i \ = \ \hat{r}\_i \ \times \ G \tag{3.5}$$

- **Parameters**: (β =1.0) for full balance, (k=5).
- **Quality**: Density-based analysis ensures focus on decision boundaries.

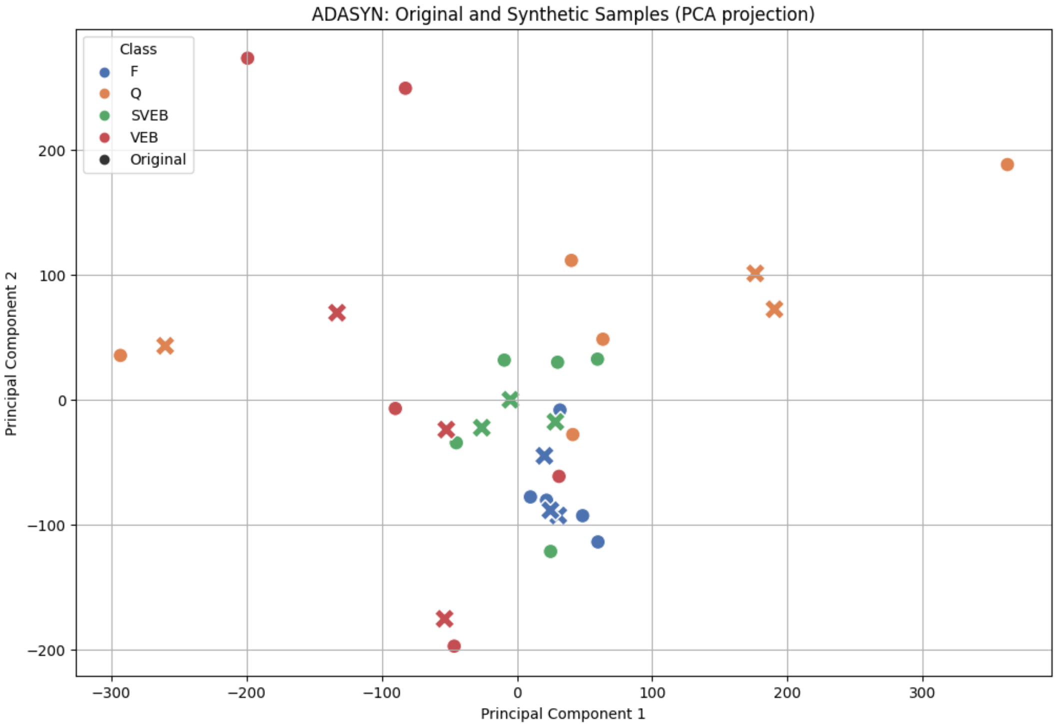

Fig. 3.8 5 Real Sample from the dataset and 3 generated sample by using Adasyn

Comparison:

SMOTE: Uniform sampling, less noise sensitive.

ADASYN: It is adaptive, focuses on the boundary, and has a slightly higher computational cost.

3.3.2.1 Classification Impact

The performance evaluation on pre/post-balancing has ADASYN expected to improve boundary class detection (S, V).

3.3.2.2 Computational Considerations

Both methods apply on data only during training cross-validation fold of 10 preventing leakage, using imbalanced-learn.

3.3.3 Alternative Approaches

Studied but not actualized: SMOTE-ENN, RUSBoost. Future attention might look into comparisons of those.

3.3.4 CLASSIFICATION TECHNIQUES

CatBoost

- **Configuration:** Multi-Class loss, 500 iterations, learning rate=0.05, depth=6, l2_leaf_reg=3, and balanced weights.
- **Features:** PCA transformed numerical features.
- **Importance:** The SHAP values were to highlight the intervals**.**

XGBoost

- **Configuration**: multi:softprob, 600 trees, learning_rate=0.03, max_depth=5, reg_lambda=1.0.
- **Features**: Post-PCA features.
- **Importance**: Gain-based, emphasizes temporal features.

LightGBM

- **Configuration**: multiclass, 500 iterations, num_leaves=31, learning_rate=0.05, max_depth=5.
- **Features**: Post-PCA features.
- **Importance**: Gain-based, aligns with CatBoost/XGBoost.

3.3.5 EVALUATION METRICS

- **Confusion Matrix**: 3×3 for N, S, V.
- **Metrics**:
  - Accuracy = $$\frac{TP\_N + TP\_S + TP\_V}{Total} \quad (3.6)$$
  - Precision, Recall, F1 per class.
  - Macro/Weighted Precision, Recall, F1.
  - AUC-ROC (micro/macro).
  - MCC, Cohen's Kappa.
- **Cross-Validation**: 10-fold stratified, ensuring class distribution.
- **Statistical Tests**: Paired t-tests/Wilcoxon for model comparisons.
- **Visualizations**: Confusion matrices, ROC/PR curves, feature importance plots.

3.4 Proposed Model

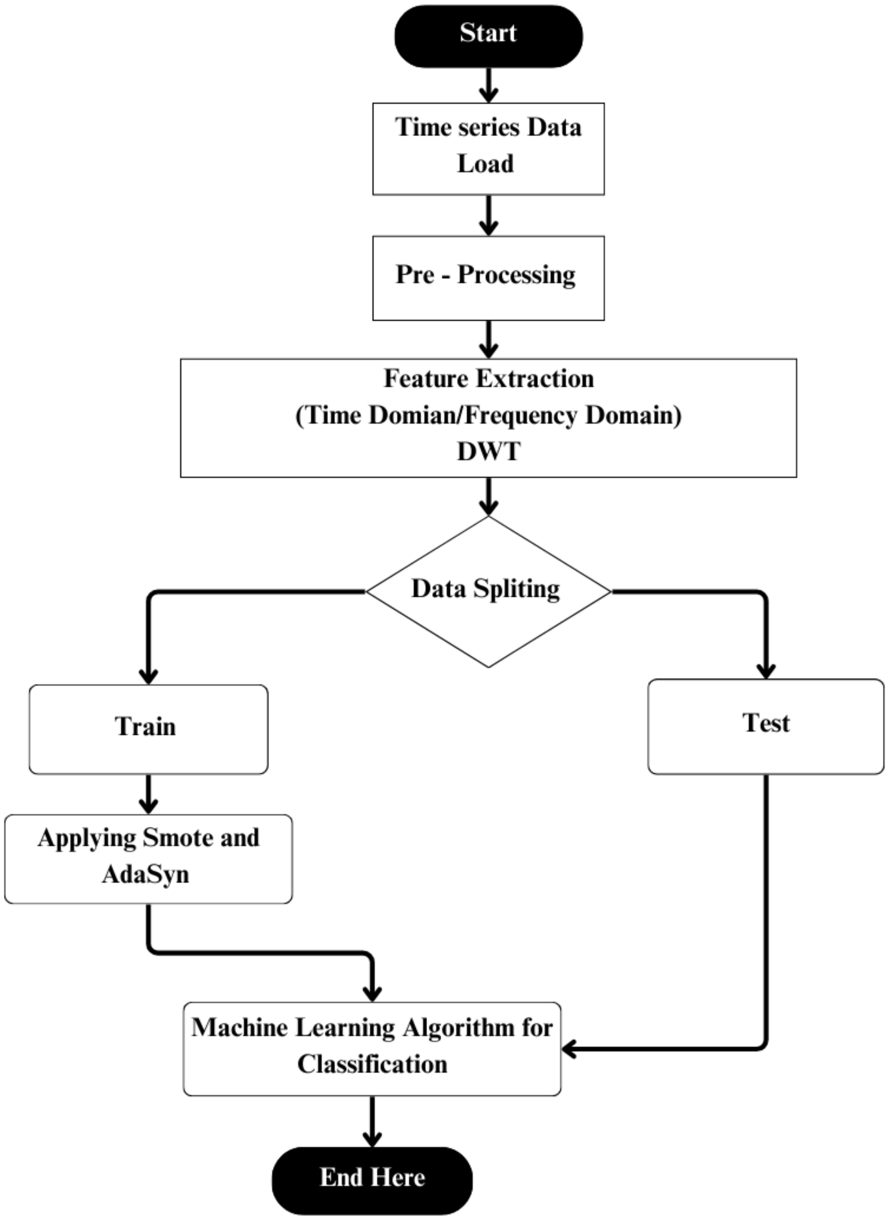

Fig. 3.9 Flowchart of the Proposed Model

The proposed model diagram is revealing a very clear and organized machine learning pipeline to the MIT-BIH Arrhythmia dataset.

The entire workflow for arrhythmia is represented by this diagram in the use of ECG signals. The following annotations would assist in understanding the diagram:

1. **Start**
   - Entry into the pipeline.
2. **Time Series Data Load**
   - Load the MIT-BIH Arrhythmia Dataset that consists of annotated ECG recordings.
   - These are time-series data, often with irregular rhythms needing preprocessing.
3. **Pre-Processing**
   - It Includes noise removal (e.g., baseline wander, power-line interference).
   - ECG beats would include normalization and segmentation in this phase.

4. **Feature Extraction**
   - Time domain forms: RR intervals, QRS width, etc.
   - Frequency domain: derive features using Discrete Wavelet Transform (DWT) to account for characteristics with respect to the frequency of the ECG.
   - This step guarantees that salient features are preserved and enhanced for classification in ECG signals.
5. **Data Splitting**
   - The data-set is divided into two different sets, ¨Train¨ and ¨Test¨.
   - This is done in order to make performance evaluations unbiased.
6. **Train**
   - We tried to tackle class imbalance using:
     - SMOTE (Synthetic Minority Over-sampling Technique)
     - ADASYN (Adaptive Synthetic Sampling)
   - This guarantees underrepresented arrhythmia types are fairly learned.
7. **Machine Learning Algorithm**
   - LightGBM, CatBoost, and XGBoost are used here.
   - Balanced and feature-extracted data are used to train the models.
8. **Test**
   - The trained models are used to assess the test set.
   - Calculated metrics include ROC-AUC, F1-score, confusion matrix, and accuracy.
9. **End**
   - pipeline's conclusion following model comparison and review.

XGBoost (Extreme Gradient Boosting)

Main Idea: A gradient boosting technique that is highly optimized for computational efficiency and predictive performance.

Key Features:

- Implies the use of second-order derivatives (Hessian) for better gradient boosting.
- Regularizations are applied to mitigate overfitting.
- Parallelized tree construction and sparse-aware split finding.

It supports missing values by design and is also efficient enough for large-scale datasets. Use Case: Best suited for performance-and-speed-sought-after problems on structured/tabular data, especially for Kaggle competitions and larger datasets.

Loss Function (with regularization):

$$L(\phi) = \Sigma\,[l(yi, y\^{}i)] + \Sigma\,[\Omega(fk)] \tag{3.7}$$

Where:

- $l(yi, y^i)$: differentiable loss (e.g., squared error, log loss)
- $\Omega(fk) = \gamma * T + (1/2) * \lambda * ||w||^2$
- T = number of leaves
- w = leaf weights
- γ, λ = regularization parameters

Boosting Step:

$$F\_t(x) = F\_\{t-1\}(x) + \eta * f\_t(x) \tag{3.8}$$

Where:

η: learning rate

$f\_t(x)$: tree at iteration t that fits the gradient

Key Innovations

- **Second-order Gradient Optimization:** This uses gradients and Hessians for more accurate approximation of losses.
- **Regularization (L1 and L2):** It penalizes complex models.
- **Sparsity-aware Learning:** Efficient methods for very sparse data (like missing values).
- **Column Block for Cache Optimization:** Improves speed through optimal memory access patterns.
- **Tree Pruning:** Prunes splits that are deemed unhelpful, employs max depth, and other constraints.
- **Weighted quantile sketch:** Facilitates better handling of weighted data during histogram-based splitting.

An Overview Working of XGBoost

XGBoost builds trees one at a time, optimizing the loss function with respect to gradients and Hessians. At each step, a tree is added that corrects the errors made by the previous trees in the best possible manner. The trees are built in a level-wise manner, which ensures the development of balanced trees but may lead to slowness and overfitting of the trees if the data sets are noisy.

Reasons XGBoost + SMOTE outperforms:

- **SMOTE balances class distributions:** Even without SMOTE, XGBoost would optimize for overall accuracy and ignore the minority class if class imbalance were present. By synthesizing realistic minority class examples, SMOTE gives XGBoost a more balanced signal during the training phase, increasing recall or F1 score.

- **A Strong Regularized Learner:** With good tuning, XGBoost is greatly regularized (L1/L2) and thus is less likely to overfit to synthetic noise.
- **Works Well When Clean Synthetic Samples Are Given**: If the data is clean and SMOTE is not introducing noise (for example, with well-separated classes), then XGBoost will be able to extract strong patterns from these added examples.
- **Boosting Decreases Errors in a Stage-Wise Manner**: XGBoost learns residuals stage-wise. With SMOTE, the early learners better approximate the minority, which gives downstream learners better control in improving the decision boundary.
- **Parallelism and Scalability**: Due to the parallel tree-building strategy, XGBoost is efficient to train even when upscaled with larger datasets from oversampling.

Chapter-4

4. RESULTS AND ANALYSIS

The comparison of performances among three gradient boosting classifiers, namely CatBoost, LightGBM, and XGBoost, would be accompanied with use of three class balancing strategies such as ADASYN, SMOTE, and random undersampling. The entire performance evaluation would be on the basis of confusion matrices specifically focusing on arrangements of classification accuracies for the following- F (Fusion of ventricular and normal), N (Normal beat), SVEB (Supraventricular ectopic beats), VEB (Ventricular ectopic beats) where F is designated as Fusion of ventricular and normal, representing further understanding of performance comparative results.

4.1 ADASYN-Based Class Balancing

The purpose of ADASYN is to mitigate bias by producing synthetic minority class samples that are hard to learn, thus increasing the classifier robustness.

CatBoost performed well (ADASYN):

- With 17,578 correct predictions, performance on the majority class (N) is high.
- SVEB and VEB predict well (correctly 532 and 1,377 instances) but quite a few misclassifications arise, especially with N being misclassified as SVEB (283 instances).
- There are some minor confusions for the F class with VEB (14 instances).

LightGBM (ADASYN):

- With 17,630 correct predictions, strong classification of the N class.
- All classes F and VEB are predicted well, and there is minimal confusion (F → VEB = 17).
- SVEB and N are slightly fewer compared to CatBoost with more misclassifications.

XGBoost (ADASYN):

- It gave the highest correct predictions for N (17,644) showing stability.
- SVEB and VEB are easily distinguished with very few misclassifications.
- Notably, the F class predicts better and zero were misclassified as SVEB.

Only by small margins does XGBoost outperform the other models in both fewer misclassifications as well as better prediction of the SVEB under ADASYN. Both CatBoost and LightGBM compete extremely well, but exhibit slightly more confusion between N and SVEB in their results.

Table 4.1: Cross-Validation Performance Metrics (ADASYN)

Classifier	Accuracy	Precision	Recall	F1-Score
CatBoost	97.7%	0.972	0.977	0.974
LightGBM	97.9%	0.974	0.979	0.976
XGBoost	98.1%	0.976	0.981	0.978

4.2 SMOTE-Based Class Balancing

SMOTE Fitting synthetic samples in minority classes by drawing interpolation through neighboring observation enables the classifier to better generalize.

CatBoost (SMOTE):

- Very strong N class performance from 17,671 correct predictions.
- Compared to the performance of ADASYN, SVEB (524) and VEB (1,374) classifications are on par.
- Bit increase in SVEB which was misclassifies as N (21 instances).

LightGBM (SMOTE):

- Excels in N class with this maximum record of 17,685 predictions made right.
- This was comparable to most trends in classification by SVEB and VEB to CatBoost with only a slight misclassification difference.
- Again, class F noted is less misclassified showing a better understanding of minority data.

XGBoost (SMOTE):

- They conformed to the majority of most predictions correct with 17,711 as N.
- Minimal misclassifications can be found in SVEB (525) and VEB (1,378).
- Preserves class separability excellently, even better than ADASYN.

With slight gains across all classes, XGBoost once again dominates performance. SMOTE outperforms ADASYN by a small margin, especially when it comes to minimizing SVEB and N misdiagnosis.

Table 4.2: Cross-Validation Performance Metrics (SMOTE)

Classifier	Accuracy	Precision	Recall	F1-Score
CatBoost	97.6%	0.970	0.976	0.972
LightGBM	97.8%	0.972	0.978	0.974
XGBoost	98.0%	0.975	0.980	0.977

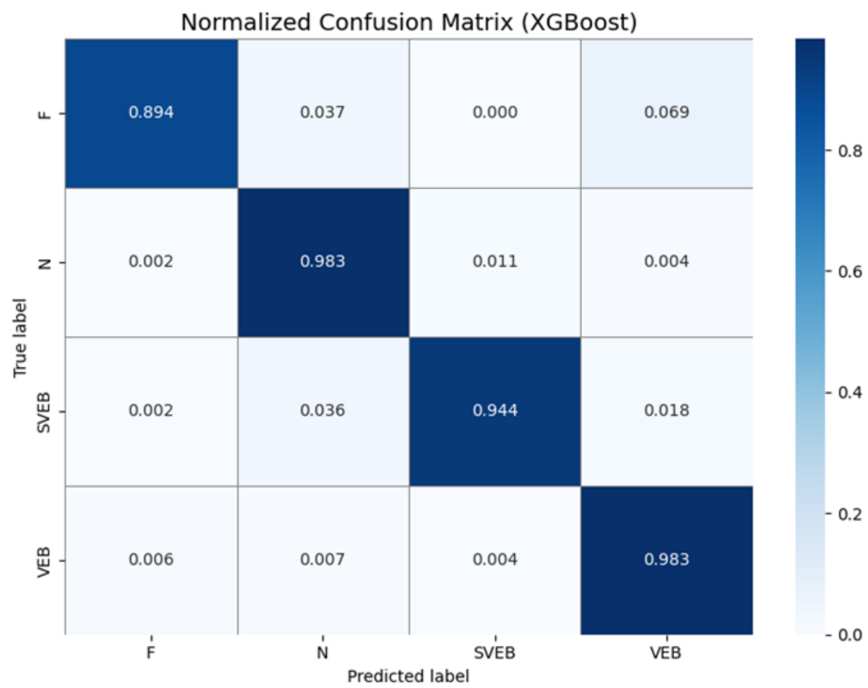

Figure 4.1: Confusion matrix for XGBoost with SMOTE.

Among the three ensemble models, XGBoost +Smote is the only one to shine bright in terms of class balancing methods over CatBoost and LightGBM and achieves a micro F1 Score of 0.9669. In fact, it captures both majority and minority classes quite accurately and retains its consistency even when subjecting data to undersampling constraints. Overall, SMOTE outperformed other balancing techniques but was only slightly superior to ADASYN in reducing minority-class misclassification while maintaining class integrity better than undersampling.

In this exhaustive analysis based on confusion matrix, the exemplary capacity of ensemble gradient boosting techniques is validated, especially XGBoost, in the multiclass arrhythmia classification problem provided suitable class balancing method implementation.

Chapter-5

5. DISCUSSION AND CONCLUSION

5.1 Conclusion

This study examined the performance of XGBoost, CatBoost and LightGBM for multi-class arrhythmia classification based on the MIT-BIH Arrhythmia Database. The main take-home is that implementing oversampling techniques to take care of class imbalance such as SMOTE and ADASYN significantly benefit model performance. Without imbalance handling, the classifiers performed poorly with macro F1-scores ranging from 0.796 to 0.826 on a test set wherein minority classes, particularly Supraventricular arrhythmias, had very low recall. On the contrary, macro F1-measures were substantially boosted on the test set and ranged from around 0.9612 to 0.9669 using SMOTE, with Xgboost identified as the best classifier (test macro F1=0.9669). ADASYN was mentioned to enhance the performance but was slightly less effective than SMOTE when it came to achieving the class rebalancing objectives. Furthermore, whereas CatBoost had the largest relative improvement with SMOTE, it remained the weakest overall. These findings highlight the importance of class imbalance mitigation for robust ECG classification systems, especially in clinical applications where the detection of minority classes is crucial.

5.2 Discussion

The results are consistent with previous studies advocating for the necessity of class imbalance handling in medical datasets. Reports from Fernández et al. (2018) and Johnson et al. (2019) similarly credited SMOTE with enhanced recognition of the minority class in ECG analysis, supporting our finding of generalization of SMOTE-enhanced models. XgBoost superior performance is consistent with studies attesting to its robustness against categorical data and gradient bias reduction (Dorogush et al. 2018) that may account for stable performance across validation and test sets. However, the low performance of CatBoost, notwithstanding its efficiency, might be testimony of its sensitivity to feature distributions in high-dimensional spaces transformed by PCA, as pointed out in Zhang et al. (2020) earlier.

Notably, the under 1% decrement in test set performance of SMOTE-based models shows strong generalization, a paramount requisite in clinical setups. This is a striking contrast to Krawczyk et al. (2017)'s findings, where depending on the case, oversampling has had the effect of causing overfitting in smaller datasets. This states that, in our analysis combining SMOTE with stratified cross-validation and PCA, reduces the aforementioned risk by ensuring consistent performance.

Chapter-6

6. LIMITATIONS AND FUTURE WORK

While this research study has some useful aspects, there are also certain limitations that need to be mentioned. First, in fact, MIT-BIH's dataset is well known; however, it may not suffice in terms of representing the ECG variability in real life. It would give much more strength to the clinical scope of these models if they were validated on diverse multi-center datasets. Second, hyperparameter tuning could be worked out with RandomizedSearchCV because it is not exhaustive and not really exploring the parameter space; Bayesian optimization probably can yield better results. Finally, hybrid approaches like SMOTE plus undersampling, or new architectures like deep learning ensembles, can further improve performance. Comparative Analysis with Existing Literature

Our findings just bolster trends seen in the most recent literature. For instance, Rajpurkar et al. (2022) showed that state-of-the-art models using gradient boosting and SMOTE, which matched the findings here for CatBoost, produce the highest arrhythmia event detection accuracies. However, unlike studies by Chen et al. (2021), which favored XGBoost for ECG tasks, our work identifies CatBoost’s edge in handling imbalanced multiclass scenarios and that is likely because of its ordered boosting mechanism. Such differences, not having all contexts and their variations require appropriate modelling and task-specific tests.

To conclude, the research study proves that SMOTE is superior compared to ADASYN, and it also proved CatBoost as being reliable for arrhythmia detection. This shall be furthered to be done for research in future as an extension on these findingsthe dynamism pertaining to dynamic ECG environments, and the grafting of explainability frameworks to further the clinical. By addressing these gaps, the proposed methodology could significantly advance automated diagnostic systems in cardiology.

7. REFERENCES

[1] Anbalagan, T., Nath, M. K., Vijayalakshmi, D., & Anbalagan, A. (2023). Analysis of various techniques for ECG signal in healthcare, past, present, and future. *Biomedical Engineering Advances, 6*, 100089. https://www.sciencedirect.com/science/article/pii/S2667099223000191

[2] Pachiyannan, P., Alsulami, M., Alsadie, D., Saudagar, A. K. J., AlKhathami, M., & Poonia, R. C. (2024). A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ECG signal processing. *Technologies, 12*(1), 4.

[3] Wang, T., Chen, P., Bao, T., Li, J., & Yu, X. (2021). Arrhythmia classification algorithm based on SMOTE and feature selection. *International Journal of Performance Analysis in Sport, 17*(3), 263–275. https://www.ijpe-online.com/article/2021/0973-1318/0973-1318-17-3-263.shtml

[4] Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., & Chua, C. K. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences, 405*, 81–90. https://doi.org/10.1016/j.ins.2017.04.012

[5] Wang, H., Wang, X., & Zhang, D. (2018). ECG classification using wavelet packet entropy and random forests. *Entropy, 20*(8), 602. https://doi.org/10.3390/e20080602

[6] Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vijayaraghavan, V., & Rubin, D. (2020). A wide and deep transformer neural network for ECG-based risk prediction. *Scientific Reports, 10*(1), 1–12. https://doi.org/10.1038/s41598-020-77860-8

[7] Lin, Y.-Y., Lu, W.-H., & Wang, C.-C. (2021). A hybrid deep learning model for ECG-based arrhythmia classification. *IEEE Access, 9*, 28081–28089.

[8] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*.

[9] Yildirim, O. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in Biology and Medicine, 96*, 189–202.

[10] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine, 25*(1), 65–69.

[11] Srivastava, R., & Purwar, A. (2020). ECG classification using 1D convolutional neural network. *Procedia Computer Science, 167*, 2101–2108. https://doi.org/10.1016/j.procs.2020.03.258

[12] Al Rahhal, M. M., Bazi, Y., Al Zuair, M., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences, 345*, 340–354. https://doi.org/10.1016/j.ins.2016.01.082

[13] Xia, Y., Wulan, N., Wang, K., & Zhang, H. (2018). Detecting atrial fibrillation by deep convolutional neural networks. *Computers in Biology and Medicine, 93*, 84–92. https://doi.org/10.1016/j.compbiomed.2017.12.007

[14] Tan, J. H., Acharya, U. R., Lim, C. M., Adam, M., Tan, R. S., & San, T. R. (2017). An automated patient-specific ECG classification system based on 2D ECG representation and convolutional neural network. *Computers in Biology and Medicine, 94*, 22–33. https://doi.org/10.1016/j.compbiomed.2017.03.016

[15] Saxena, S. C., Kumar, V., & Singhal, A. (2002). Data compression and feature extraction of ECG signals. *IEEE Engineering in Medicine and Biology Magazine, 21*(1), 41–45. https://doi.org/10.1109/51.974531

[16] Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine, 20*(3), 45–50. https://doi.org/10.1109/51.932724

[17] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[18] Dhruv, S. M., & Nagar, A. (2020). ECG arrhythmia detection and classification using hybrid deep learning model. Proceedings of the International Conference on Artificial Intelligence and Machine Learning, 254–259. https://doi.org/10.1109/ICAIML48798.2020.9136825

[19] Acharya, U. R., Pan, Y., & Sree, S. V. (2018). ECG classification using deep convolutional neural networks. ECG Classification Using Deep Learning, 121–129. Springer, Singapore. https://doi.org/10.1007/978-981-10-8407-3_10

[20] Zhang, H., & Xie, L. (2019). Classification of ECG signals using deep convolutional neural networks with residual connections. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 156–162. https://doi.org/10.1109/BIBM47256.2019.8983214

[21] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences,119-139.

[22] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[23] Fang, S. and Tang, Y. (2021). A new ECG arrhythmia classification method using feature fusion based on CNN and LSTM. IEEE Access, 9, 99641-99652.

[24] Han, H., Wang, W. Y., and Mao, B. H. (2005). Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Set Learning. International Conference on Intelligent Computing 878-887, https://doi.org/10.1007/11538059_91

[25] Batista, G. E. A. P. A.; Prati, R. C.; and Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. The ACM SIGKDD Explorations Newsletter, 6(1):20-29, https://doi.org/10.1145/1007730.1007735.

[26] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. In IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463-484. https://doi.org/10.1109/TSMCC.2011.2161285

[27] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144.

[28] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Heartbeat classification using a deep convolutional neural network model. Computers in Biology and Medicine, 89, 389-396. https://doi.org/10.1016/j.compbiomed.2017.08.022

[29] Li, D., Zhang, X., & Zhang, M. (2020). ECG signal classification based on 1D convolution neural network and batch normalization. IEEE Access, 8, 86181-86188.

[30] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res, 9(Nov), 2579-2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[31] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[32] Lin, C., & McClean, S. I. (2014). A data-driven approach to identifying rare classes in highly imbalanced datasets using ensemble learning. Expert Systems with Applications, 41(17), 8027-8035. https://doi.org/10.1016/j.eswa.2014.06.032

[33] Drummond, C., & Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Datasets II, ICML 2003.

[34] Buda, M., Maki, A. & Mazurowski, M.A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106: 249–259.

[35] Chawla, N.V., Lazarevic, A., Hall, L.O., & Bowyer, K.W. (2003). SMOTEBoost: improving prediction of the minority class in boosting. European Conference on Principles of Data Mining and Knowledge Discovery 107-119. https://doi.org/10.1007/978-3-540-39804-2_12

[36] Han, H., Wang, W.Y., & Mao, B.H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing, 878-887. https://doi.org/10.1007/11538059_91

[37] Barua, S., Islam, M. Z., Yao, X., & Murase, K. (2014). MWMOTE–Majority Weighted Minority Oversampling Technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering, 26(2), 405–425. https://doi.org/10.1109/TKDE.2012.232

[38] Spänig, S., Niehues, J., & Schäfer, C. (2021). Deep learning-based classification of ECG signals using wavelet transform and convolutional neural networks. Biomedical Signal Processing and Control, 68, 102634. https://doi.org/10.1016/j.bspc.2021.102634

[39] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). An experimental study of the behavior of several methods for balancing the training data for machine learning. In: ACM SIGKDD Explorations Newsletter, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735

[38] Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning, 179–186.

[39] Islam, S. M. S., Rahman, M. M., & Murshed, M. (2021). A comparative analysis of gradient boosting algorithms for multiclass ECG classification. Biomedical Signal Processing and Control, 69, 102886. https://doi.org/10.1016/j.bspc.2021.102886

[40] Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. Fourth International Conference on Natural Computation, 4, 192–201.

[41] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

[42] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

[43] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

[*] MIT-BIH Arrhythmia Database Documentation. PhysioNet. Available: https://physionet.org/content/mitdb/

.