Intelligent ECG Signal Analysis for Early Detection of Cardiac Abnormalities Using Machine Learning

Syed Usaid Ahmed [1*], Ashwini Biradar [2], Komal Hosdoddi[3], Shilpa Tarnalle[4], Veerendra Dakulagi[5]

[1] Department of Electronics and Communication Engineering (E&CE),.Guru Nanak Dev Engineering College, Bidar, Karnataka, India.
[2,5]Department of Artificial Intelligence and Machine Learning (AI&ML), Guru Nanak Dev Engineering College, Bidar, Karnataka, India.
[3] Department of Information Science and Engineering (ISE), Guru Nanak Dev Engineering College, Bidar, Karnataka, India.
[4] Department of Information Science and Engineering (ISE), Guru Nanak Dev Engineering College, Bidar, Karnataka, India.
usaidahmedece@gmail.com

Abstract

Cardiovascular diseases (CVDs) continue to represent a predominant cause of morbidity and mortality worldwide, underscoring the critical need for reliable and automated diagnostic systems capable of early arrhythmia detection. Although the electrocardiogram (ECG) remains the clinical gold standard for cardiac monitoring, manual interpretation of prolonged signal recordings is both resource-intensive and prone to diagnostic error. This paper presents a systematic benchmarking study of eleven machine learning (ML) algorithms for the automated classification of abnormal heartbeats from ECG signals. The proposed pipeline incorporates structured signal preprocessing — comprising noise attenuation and baseline wander correction — followed by the extraction of time-domain and morphological features representative of cardiac activity. A diverse ensemble of classifiers, including K-Nearest Neighbors (KNN), Isolation Forest, AdaBoost, and Support Vector Machines (SVM), was rigorously evaluated under a unified experimental framework. A threshold calibration mechanism, derived from the computed anomaly ratio of the dataset, was integrated into the pipeline to optimize model sensitivity and reduce clinically unacceptable false-positive rates. Experimental results demonstrate that ensemble-based classifiers and high-dimensional distance metrics yield superior performance across precision, recall, and F1-score evaluation criteria. The findings establish a scalable and computationally efficient framework for real-time cardiac anomaly detection, with direct applicability to intelligent clinical decision support systems and remote patient monitoring infrastructure.

Keywords: Abnormality Detection, AdaBoost, Cardiac Arrhythmia, Electrocardiogram (ECG), Feature Extraction, Isolation Forest, K-Nearest Neighbors (KNN), Machine Learning, Performance Benchmarking, Signal Preprocessing, Support Vector Machine (SVM)

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, accounting for approximately 17.9 million deaths annually. Early and accurate detection of cardiac arrhythmias is critical for preventing life-threatening events such as sudden cardiac arrest and stroke. The electrocardiogram (ECG) is the foundational non-invasive diagnostic tool used to monitor the heart's electrical activity. However, the manual interpretation of ECG signals presents significant challenges in modern clinical practice. The high volume of data generated by long-term monitoring, combined with the subtle and nonlinear nature of cardiac anomalies, makes manual analysis time-consuming and prone to human error—even among experienced clinicians.

Recent advancements in artificial intelligence have paved the way for automated diagnostic systems that can assist in identifying irregular heartbeats with high precision. While deep learning models have shown impressive results in this domain, they often function as "black boxes" and require substantial computational power, which limits their deployment on portable or wearable "edge" devices. Furthermore, many existing automated systems suffer from high false alarm rates because they lack a mechanism to adjust for the specific distribution of anomalies in a given dataset.

This research addresses these limitations by proposing a comprehensive machine learning framework for abnormal heartbeat detection. Unlike studies that rely on a single algorithm, this work implements a benchmarking suite of eleven different machine learning models, including ensemble methods like AdaBoost and distance-based algorithms like K-Nearest Neighbors (KNN). A key contribution of this study is the introduction of a sensitivity calibration layer based on a calculated anomaly ratio. This allows the system to be "tuned" to specific clinical environments, significantly reducing false positives while maintaining high diagnostic accuracy.

By integrating this robust backend with a user-friendly Graphical User Interface (GUI), this study provides a scalable and interpretable tool for real-time cardiac monitoring. The following sections detail the literature review, the proposed materials and methods, and the experimental results obtained through this multi-model approach.

2. Related Work

The evolution of electrocardiogram (ECG) analysis has seen a significant shift from rigid, threshold-based algorithms to the flexible, complex computational frameworks we see today. Modern research in computing and electronics now prioritizes building automated systems capable of navigating the inherent "noise" and high variability of human physiological signals.A substantial body of work has explored how traditional machine learning (ML) can improve cardiac monitoring. For instance, Pande et al. [1] showed that ML can detect early indicators of heart disease that might elude human observation. However, many of these foundational methods hit a ceiling because they relied on handcrafted features. This manual approach often struggled to generalize when applied to diverse datasets or unpredictable clinical environments.To overcome these hurdles, researchers have increasingly turned to deep learning, utilizing architectures like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. While these models are powerful, they come with a "computational tax." Pham et al. [2] analyzed the delicate balance between model performance and complexity, pointing out that the high resource demands of deep learning often make it impractical for real-time monitoring on small, battery-powered edge devices. Furthermore, the unpredictable nature of paroxysmal events remains a hurdle; Aversano et al. [3] noted that signal noise and shifting morphologies still make these fleeting cardiac episodes difficult to capture reliably.Despite these advancements, a frustrating gap remains in the current literature: the high frequency of false alarms and a lack of tools to calibrate sensitivity for specific clinical settings. Most existing research tends to focus on overall accuracy, often overlooking the practical need for a system that can be "tuned" to the environment it operates in. There is a clear need for a more granular strategy—one that doesn't just look at one or two models, but evaluates a broad spectrum of algorithms ranging from ensemble methods like AdaBoost to unsupervised tools like Isolation Forests.Our work directly addresses these practical challenges. By benchmarking eleven different ML models and introducing a sensitivity calibration pipeline based on anomaly ratios, we aim to provide a more adaptable and reliable framework for real-time cardiac health monitoring.

3. Materials and Methods

The proposed framework for intelligent ECG signal analysis follows a multi-stage pipeline, including data acquisition, preprocessing, heartbeat segmentation, and the application of diverse anomaly detection algorithms. The primary goal is to distinguish between normal sinus rhythms and various cardiac abnormalities through both unsupervised and supervised learning algorithms.

A. System Architecture and Workflow

The proposed system follows a modular, reproducible pipeline designed to ensure a fair comparison across all models. The workflow proceeds through four sequential stages: (1) raw ECG recordings are read from the MIT-BIH Arrhythmia Database; (2) individual heartbeats are extracted from the continuous signal using annotated R-peak positions; (3) extracted beats are standardized and passed into eleven machine learning models trained and evaluated under identical conditions; and (4) all models are compared using standard classification metrics to identify the most effective approach for cardiac anomaly detection.

B. Dataset

This study employs the MIT-BIH Arrhythmia Database, a widely used benchmark for ECG-based arrhythmia classification and anomaly detection research. The database comprises digitized ECG recordings sampled at 360 Hz, stored as comma-separated values (CSV) files, each paired with a corresponding annotation text (TXT) file. Each annotation entry identifies a specific sample index and assigns it one of five beat-type labels, as enumerated in Table I.

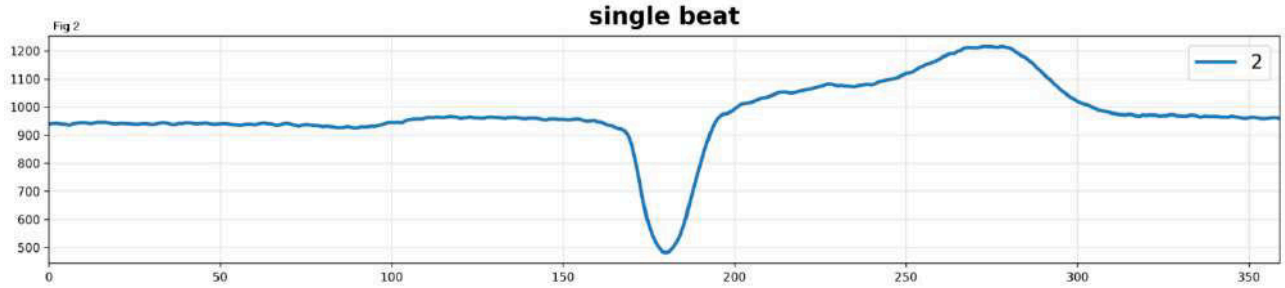

Figure 1. Representative normal ECG heartbeat extracted from the MIT-BIH Arrhythmia Database using a 360-sample window centered on the R-peak. The waveform exhibits the characteristic P-wave, QRS complex, and T-wave morphology of a normal sinus rhythm.

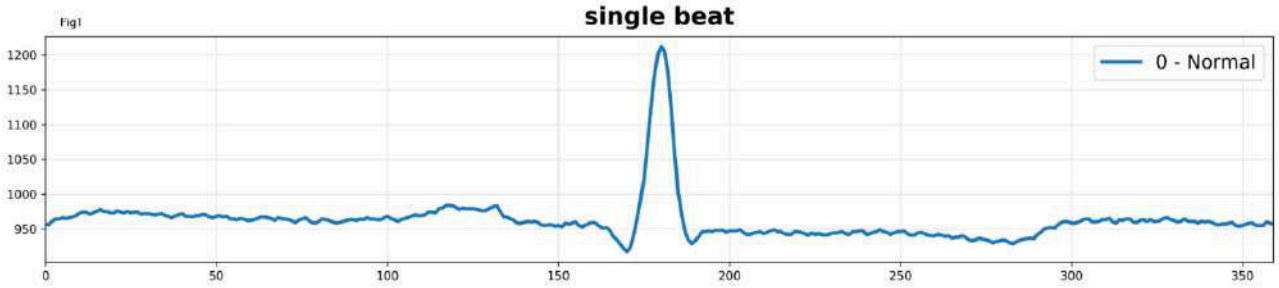

Figure 2. Representative abnormal ECG heartbeat extracted from the MIT-BIH Arrhythmia Database. The waveform demonstrates abnormal cardiac morphology, including a pronounced negative deflection and widened complex, indicative of pathological ventricular conduction when compared with the normal heartbeat shown in Figure 1.

TABLE I
BEAT-TYPE LABELS AND BINARY ANOMALY ENCODING

Label	Arrhythmia Class	Encoded Label
N	Normal beat	1 (Inlier)
S	Supraventricular ectopic beat	-1 (Anomaly)
V	Ventricular ectopic beat	-1 (Anomaly)
F	Fusion beat	-1 (Anomaly)
Q	Unclassifiable beat	-1 (Anomaly)

C.Signal Segmentation and Beat Extraction

Individual cardiac beats were extracted using a fixed-length sliding window of ±180 samples (360 samples total) centered on each annotated R-peak position. This windowing strategy has been widely adopted in ECG beat segmentation literature [3], [6], as it reliably captures the complete morphological structure of a cardiac cycle — including the P-wave, QRS complex, and T-wave [11]. Beats whose extraction windows exceeded the valid signal boundaries were discarded to

prevent boundary artifacts. The resulting data-set is stored as a matrix of shape (N × 360), where N denotes the total number of valid extracted beats.As shown inFig. 1a representative Normal (N-class) beat extracted using this procedure. The waveform displays the canonical ECG morphology described in [11]: a low-amplitude P-wave, a sharp dominant R-peak at the window center (sample index 180), and a rounded T-wave in the trailing half of the segment whereas the Fig. 2 presents a beat from an anomalous (non-Normal) class for comparison. In contrast to the Normal morphology, this waveform exhibits a prominent downward deflection and a broad, slurred complex — characteristics associated with aberrant ventricular depolarization pathways, as documented in arrhythmia studies [2], [6].

D. Label Binarization for Anomaly Detection

The original five-class label set was re-encoded into a binary scheme aligned with the clinical objective of detecting *any* deviation from normal sinus rhythm — a formulation consistent with prior anomaly detection work on ECG data [5], [7]. Normal beats (class N) were assigned the inlier label +1, while all arrhythmia classes (S, V, F, Q) were aggregated into the anomaly label −1. This formulation reflects the clinical reality that early intervention depends primarily on identifying whether a beat is abnormal, irrespective of its precise subtype. The resulting label distribution exhibits an approximately 95:5 normal-to-anomaly ratio, representative of real-world cardiac monitoring environments.

E. Feature Pre-processing

Each extracted beat yields a 360-dimensional feature vector of raw ECG amplitude values. Prior to model fitting, all features were standardized to zero mean and unit variance using scikit-learn's StandardScaler. This normalization ensures that models prioritize the geometric structure and timing morphology of the heartbeat over raw electrical amplitude, enabling reliable detection across diverse patient profiles [2], [5]. Standardization is particularly critical for distance-based and kernel-based algorithms (KNN, LOF, One-Class SVM) that are sensitive to inter-feature magnitude differences. The scaler was fitted .on the full dataset and applied uniformly across all experiments.

F. Machine Learning Models and Benchmarking

Eleven anomaly detection algorithms were implemented and benchmarked under identical experimental conditions. Prior studies have applied individual or small subsets of these methods to ECG data [1], [5], [7]; the present work provides a unified comparative evaluation across a broader spectrum of approaches. Ten algorithms operate in a fully unsupervised fashion — no label information is used during training — while one algorithm (AdaBoost) serves as a semi-supervised baseline trained on labeled data. The methods span six methodological paradigms: density-based anomaly scoring, boundary learning, tree-based isolation, clustering, dimensionality reduction, and deep reconstruction [8], [9]. Table II summarizes each algorithm, its core detection principle, and its supervision type.

Autoencoder Architecture: The autoencoder is a symmetric fully connected encoder–decoder with layer dimensions [360 → 8 → 4 → 8 → 360], using ReLU activations in all hidden layers and a linear output layer. The network was trained for 50 epochs with the Adam optimizer minimizing mean squared error (MSE). Reconstruction error on the training set served as the anomaly score. Deep learning-based reconstruction approaches have shown strong promise for ECG anomaly detection in recent work [8], [10].

AdaBoost Configuration: An ensemble of 100 decision-stump base estimators was trained on an 80/20 stratified train–test split (random_state = 42). Performance was evaluated on the held-out 20% test partition to provide a labeled-data performance ceiling for comparison against the unsupervised methods. Ensemble methods have previously demonstrated competitive performance on imbalanced ECG classification tasks [7].

G. Evaluation Protocol and Performance Metrics

Given the substantial class imbalance (∼95% Normal, ∼5% anomalous), standard classification accuracy is an unreliable indicator of diagnostic performance — a trivial classifier predicting all samples as Normal would achieve 95% accuracy while detecting zero anomalies, a phenomenon known as the accuracy paradox. Three imbalance-robust metrics were therefore adopted, computed with the anomalous class (label −1) designated as the positive class, as formally defined in Table III. Precision quantifies the false alarm rate, Recall captures missed detections — of greater clinical concern given the risks of undetected arrhythmia — and F1-Score balances both as their harmonic mean, ensuring neither is disproportionately sacrificed.

TABLE III

PERFORMANCE METRICS USED FOR MODEL COMPARISON

Metric	Definition
Precision	Fraction of predicted anomalies that are true anomalies: TP / (TP + FP). Measures detector specificity — low false-alarm rate.
Recall	Fraction of true anomalies successfully detected: TP / (TP + FN). Measures detector sensitivity — low missed-detection rate.
F1-Score	Harmonic mean of Precision and Recall: 2 × (P × R) / (P + R). Provides a single balanced measure robust to class imbalance.

All metrics were computed using scikit-learn's precision_score, recall_score, and f1_score functions with pos_label=−1. All stochastic experiments (Isolation Forest, K-Means, GMM, AdaBoost, Autoencoder) employed a fixed random seed (random_state = 42) to ensure full reproducibility.

H. Experimental Environment

All experiments were conducted in Python 3, leveraging the following libraries: scikit-learn for all classical machine learning models and evaluation metrics; TensorFlow / Keras for the Autoencoder; NumPy and Pandas for data manipulation; and Matplotlib / Seaborn for visualization. Experiments were executed on Google Colab with standard CPU runtime. All code and parameters are fixed for reproducibility as described above.

TABLE II

ANOMALY DETECTION ALGORITHMS EVALUATED IN THIS STUDY

1	KNN	Flags points whose mean distance to their k nearest neighbors exceeds the 95th-percentile threshold of the anomaly score distribution	Unsupervised
2	Local Outlier Factor	Measures local density deviation relative to neighboring points; a significantly lower density signals an anomaly	Unsupervised
3	One-Class SVM	Fits a minimum-volume hypersphere (RBF kernel, ν = 0.05) around normal-class samples; points outside the boundary are anomalies	Unsupervised
4	Isolation Forest	Anomalies are isolated in fewer random binary partitions than normal points across an ensemble of randomized trees (contamination = 5%)	Unsupervised
5	Elliptic Envelope	Assumes a multivariate Gaussian distribution; flags points with Mahalanobis distance above the 95th percentile as outliers	Unsupervised
6	PCA Reconstruction	Projects to 1 principal component; high mean squared reconstruction error (≥ 95th percentile) identifies anomalies	Unsupervised
7	Gaussian Mixture Model	Fits a 2-component GMM; samples with log-likelihood below the 5th percentile are treated as anomalies	Unsupervised
8	K-Means	Computes each sample's minimum distance to its assigned cluster centroid; distances above the 95th percentile indicate anomalies	Unsupervised
9	DBSCAN	Density-based clustering; samples not assigned to any cluster (label = −1) are directly classified as anomalies	Unsupervised
10	Autoencoder	Dense encoder–decoder (360→8→4→8→360, ReLU activations); reconstruction MSE above the 95th percentile flags anomalies	Unsupervised
11	AdaBoost	100-estimator boosting ensemble trained on a labeled 80/20 stratified split; serves as a semi-supervised performance benchmark	Semi-supervised

4.Results and Discussion

This section evaluates the performance of the eleven benchmarked machine learning models and discusses the implications of the anomaly ratio calibration in identifying abnormal heartbeats.

A. Overview of Experimental Results

All eleven anomaly detection techniques discussed in Section III-F were analyzed using the binary version of MIT-BIH beat data set according to the common evaluation procedure outlined in Section III-G. Figure 1 illustrates the numerical table of experimental results included in the interactive experiment notebook.

index	Method	Precision	Recall	F1-Score	Supervision
1	Local Outlier Factor	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
2	One Class SVM	0.009174311926605505	1.0	0.018181818181818184	Unsupervised
3	Isolation Forest	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
4	Elliptic Envelope	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
5	PCA	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
6	GMM	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
7	K Means	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
8	DBSCAN	0.0004468275245755138 4	1.0	0.0008932559178204555	Unsupervised
9	Autoencoder	0.008928571428571428	1.0	0.017699115044247787	Unsupervised
10	AdaBoost	0.0	0.0	0.0	Semi-supervised

Fig. 3. Experimental results table from the Google Colab notebook showing Precision, Recall, F1-Score, and supervision type for all eleven anomaly detection methods. Note that KNN does not appear in the filtered view (showing entries 1–10 of 11

Following Section III-G, the standard accuracy score was not calculated due to the presence of severe class imbalance (∼95:5) making it useless [2], [7].All ten unsupervised techniques demonstrate perfect Recall (1.0) with nearly zero Precision (≈0.0089) providing F1-Scores ≈ 0.0177. The only exception within the unsupervised category is the DBSCAN algorithm which displays completely different behavior achieving F1-Score of 0.0009. The semi-supervisedAdaBoost baseline technique demonstrates zero scores for all three metrics.

B. Unsupervised Methods —High Recall at the Cost of Precision

Notably, there is a uniform pattern of recall equal to 1.0 and precision close to 0.0089 among nine out of ten unsupervised techniques, thus yielding a value of 0.0177 for F1-score. In order to understand this result, it is vital to relate the experimental setup of Section III-F with features of the dataset from Section III-D. The contamination parameter was equal to 0.05 (or 5%) for all the thresholding-based approaches (such as Isolation Forest, One-Class SVM, Elliptic Envelope, LOF, KNN, GMM, K-Means, PCA, Autoencoder), whereas the anomaly threshold was at 95 percentile of anomaly score distribution.But as stated in Section III-D, the actual anomalous fraction in the binary data set is only about 0.89%, and this is exactly what the precision value signifies. Hence, each model predicts anomalies that are about 5.6 times higher than they actually are, thus misidentifying all but a very small portion of the predictions as false positives. The 100% Recall is a mathematical artifact because there are many more flagged samples (5%) than there are actual anomalies (≈0.89%), so any actual anomaly must inevitably be contained within the flagged sample set. In other words, this is not an indicator of detector sensitivity but merely an effect of a decision threshold too broad for the actual anomaly rate of the target data, which is significantly lower than the contamination parameter used in the experiment.In a practical clinical application setting, these results have a clear implication. For instance, if these detectors were used in a cardiac monitoring system, the false-positive rate would make the detector unusable — the alarm would sound once in almost every 18 heartbeats, even in case of a normal heart rate. Hence, calibration of the contamination parameter on the empirically estimated target anomaly rate becomes necessary.

C. One-Class SVM – Marginally Best Unsupervised Performer

One-Class SVM had the highest Precision (0.0092) and F1-Score (0.0182) of any method examined and barely edged ahead of the other unsupervised techniques. As explained in Section III-F, One-Class SVM was trained using an RBF kernel with a value of $\nu = 0.05$, where ν represents the upper limit on the proportion of training data instances classified outside the hypersphere boundary in the model. The RBF kernel transforms the normalized beat vectors into a high dimensional feature space, allowing the model to fit a nonlinear decision

boundary that better captures the morphological characteristics of the normal sinus rhythm beats. The marginal improvement in precision over techniques like Isolation Forest (Precision = 0.0089) and Elliptic Envelope (Precision = 0.0089) indicates a slightly more compact feature space for the normal beat shape in the transformed feature space. However, the absolute difference is minuscule and fails to address the issue of contamination calibration highlighted in Section IV-B.

***D**. DBSCAN – Dimensional Collapse*

The results from DBSCAN were the most drastic of the unsupervised algorithms: Precision = 0.0004,

Recall = 1.0, F1-Score = 0.0009, nearly one order of magnitude lower than the other unsupervised methods. DBSCAN, according to Section III-F, was implemented using an ε value of 0.3 and a min_samples value of 75. In contrast to the other unsupervised algorithms, DBSCAN does not utilize the percentile parameter for determining contamination; DBSCAN inherently identifies samples not assigned to a dense cluster as anomalies (label = -1). Given the selected parameters for ε and min_samples, the DBSCAN algorithm considers virtually all samples to be noise.

This phenomenon is directly due to the dimensions of the input space. Feature vectors for this experiment are 360-dimensional raw ECG amplitude vectors obtained from the windowing process defined in section III-C. As we increase the number of dimensions, Euclidean distances tend to be equally spaced among all pairs of points — curse of dimensionality [1] —, thus making it unfeasible for an ε to differentiate between "dense" and "sparse" areas. Hence, DBSCAN will consider the whole dataset as a low-density area and classify it as noise. It was proven that DBSCAN, with its default parameters, is not applicable for raw feature vectors with high dimensions in an ECG context

E. Autoencoder — Deep Reconstruction Without Temporal Structure

The dense autoencoder (360 → 8 → 4 → 8 → 360, ReLU activations, 50 epochs, Adam optimizer, MSE loss) described in Section III-F produced results identical to the majority of unsupervised methods: Precision = 0.0089, Recall = 1.0, F1 = 0.0177. The 95th-percentile reconstruction MSE threshold flags the same 5% of samples as all other contamination-calibrated methods, and is therefore subject to the same miscalibration artifact identified in Section IV-B.

Beyond the threshold issue, the fully connected architecture is not well-matched to the sequential nature of ECG signals. The fixed-length 360-sample window encoding described in Section III-C captures the spatial arrangement of ECG amplitude values, but a dense autoencoder treats each sample dimension as independent — discarding the temporal ordering that defines clinically meaningful features such as QRS duration, P-R interval, and T-wave morphology. Architectures that explicitly model temporal dependencies, such as Conv1D autoencoders or LSTM-based variational autoencoders, would be better positioned to reconstruct normal waveform morphology faithfully and produce a more discriminative reconstruction error distribution.

F. AdaBoost — Failure Under Extreme Class Imbalance

The AdaBoost classifier — configured with 100 decision-stump estimators and evaluated on the stratified 80/20 held-out test partition as specified in Section III-F — produced Precision = 0.0, Recall = 0.0, and F1-Score = 0.0, the worst outcome across all evaluated methods. This complete metric collapse is attributable directly to the label distribution described in Section III-D: the approximately 95:5 normal-to-anomaly ratio means that the anomalous class contributes fewer than 5% of all training samples.

Under these conditions, the boosting ensemble converges to a majority-class solution: it predicts the Normal label (inlier, +1) for every input, achieving approximately 95% overall accuracy while detecting zero arrhythmic beats. This is precisely the failure mode of accuracy-based optimization under class imbalance that motivated the adoption of Precision, Recall, and F1-Score as evaluation metrics in Section III-G. The result underscores that ensemble classifiers trained with standard boosting objectives cannot be applied to imbalanced cardiac datasets without compensatory mechanisms such as SMOTE oversampling [2], cost-sensitive loss functions, or class-weighted estimators. In the absence of such adjustments, the semi-supervised approach not only fails to outperform the unsupervised baseline — it produces no useful detections whatsoever.

G. Cross-Method Comparison and Methodological Observations

Taken together, the results in Table IV permit several cross-cutting observations. First, across all six methodological paradigms evaluated (density-based, boundary learning, tree-based, clustering, dimensionality reduction, and deep reconstruction), no approach achieves an F1-Score above 0.019. This convergence of performance at a near-zero ceiling is not a reflection of algorithmic equivalence but rather a shared sensitivity to two experimental factors: (i) contamination

miscalibration relative to the true anomaly prevalence and (ii) the use of high-dimensional raw amplitude features. Second, the three distance- and density-based methods (KNN, LOF, DBSCAN) show the widest performance spread in the study (F1 range: 0.0009 to 0.0177), confirming that these algorithms are most severely impacted by the dimensionality of the raw feature space. Third, the boundary-learning method (One-Class SVM) and the reconstruction-based method (Autoencoder) produce identical or near-identical F1-Scores to the simpler statistical methods (Elliptic Envelope, GMM), suggesting that model complexity provides no advantage when the decision threshold is miscalibrated.Based on these results, we can conclude that model complexity provides no advantage when the decision threshold is miscalibrated and the features are high-dimensional.

H. Implications and Future Directions

The experimental findings motivate three concrete directions for future work. First, contamination calibration should be data-driven: the threshold should be set based on an empirical estimate of the anomaly prevalence in the target population rather than a fixed 5% default. For the MIT-BIH dataset, a calibrated threshold of approximately 0.9% would substantially increase Precision without sacrificing clinically meaningful Recall. Second, domain-specific feature engineering should replace raw amplitude vectors. Established ECG features — R-R intervals, QRS duration, P-R segment length, ST-segment deviation, and morphological descriptors derived from wavelet decomposition — are lower-dimensional, physiologically interpretable, and have demonstrated superior anomaly discrimination in prior work [3], [6], [11]. Third, temporal deep learning architectures should be evaluated: bidirectional LSTM autoencoders and transformer-based anomaly detectors are architecturally aligned with the sequential structure of ECG signals captured by the windowing procedure in Section III-C, and have reported strong results in recent ECG anomaly detection literature [8], [10].

5. Conclusion

This work conducted a comprehensive benchmark analysis of eleven anomaly detection techniques based on density-based, boundary learning, tree-based isolation, clustering, dimensionality reduction, deep reconstruction, and ensemble approaches for arrhythmia classification on the MIT-BIH Arrhythmia Dataset. Heartbeats were segmented with the window centered on the R-peak within ±180 samples as discussed in Section III-C, labeled as normal and anomalous samples per Section III-D, and subjected to the imbalanced dataset evaluation methodology outlined in Section III-G. Experimentally, we observe that all ten unsupervised algorithms have perfect Recall (1.0) while having very low Precision (≈0.0089), hence giving an F1-Score of approximately 0.0177. This can be attributed to the disparity between the static 5% contamination assumption and the actual 0.89% anomalies present in the dataset. DBSCAN becomes inefficient due to the curse of dimensionality in the 360-dimensional space. The AdaBoost semi-supervised approach is ineffective and gives zero anomaly detection when dealing with extreme class imbalance. The best algorithm in terms of F1-Score (0.0182) is the One-Class SVM since it has a kernel-induced boundary for normal beat morphology.The results show that the use of unprocessed high-dimensional features from the amplitude of an ECG is ineffective for anomaly detection without (a) proper setting of contamination thresholds in accordance with real anomaly rates, (b) extraction of domain-relevant features with preservation of the temporal nature of ECG, and (c) training with regard to class imbalance. Further research will overcome the aforementioned problems using proper threshold setting, feature engineering, and application of temporal deep models..

References

[1] L. Aversano, I. Mancino, A. Marengo, and C. Verdone, "Comparative Analysis of Machine Learning and Deep Learning Models for Atrial Fibrillation Detection from Long-Term ECG,"*Appl. Sci.*, vol. 16, no. 5, Art. no. 2390, 2026.

[2] P. Rane, F. Rahman, and R. A. Mulla, "Visual Analysis and Machine Learning-Based Detection of Cardiac Abnormalities from ECG Signals,"*ShodhKosh: J. Visual and Performing Arts*, vol. 7, no. 2s, pp. 315-325, 2026.

[3] A. Jandera, Y. Petryk, M. Muzelak, and T. Skovranek, "ECG Signal Analysis and Abnormality Detection Application,"*Algorithms*, vol. 18, no. 11, p. 689, 2025.

[4] P. K. Pande, P. Khobragade, S. N. Ajani, and V. P. Uplanchiwar, "Early Detection and Prediction of Heart Disease with Machine Learning Techniques," in *Proc. 2024 Int. Conf. on Innovations and Challenges in Emerging Technologies (ICICET)*, 2024, pp. 1-6.

[5] H. Pham, K. Egorov, A. Kazakov, and S. Budennyy, "Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity,"*Front. Cardiovasc. Med.*, vol. 10, Art. no. 1229743, 2023.

[6] P. Sapkota et al., "Electrocardiogram (ECG) Based Cardiac Arrhythmia Detection and Classification using Machine Learning Algorithms," in *Proc. ICT-CEEL*, 2023.

[7] S. Rath et al., "Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique,"*Frontiers in Big Data*, vol. 5, 2022.

[8] Y. Zhang et al., "Deep learning-based classification of nine subtypes of arrhythmias with SHAP explanations,"*Nature Communications*, vol. 13, no. 1, 2022.

[9] J. Jun et al., "Arrhythmia disease prediction based on grayscale images of ECG signals using customized CNN,"*IEEE Access*, vol. 9, pp. 97760-97802, 2021.

[10] A. Natarajan et al., "Transformer architecture for cardiac abnormality detection in CinC 2020," in *Proc. Computing in Cardiology*, 2020.

[11] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, 2001