



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ecg-Based Heart Disease Prediction

¹Sheetal. P. Patil, ²Sunny. D. Patil, ³Nikhil. V. Songire, ⁴Kunal. D. Desale, ⁵Aditya. O. Shukla

¹Shirpur, ²Shirpur, ³Shirpur, ⁴Shirpur, ⁵Shirpur

¹Artificial Intelligence And Machine Learning,

¹R. C. Patel Institute of Technology, Shirpur, India

Abstract: Cardiovascular diseases (CVDs) are the leading cause of death globally. Early prediction and detection of heart related anomalies using electrocardiogram (ECG) signals can significantly improve patient outcomes. This research proposes an ensemble-based machine learning approach for heart disease prediction using ECG data from the PhysioNet database. The system integrates multiple classifiers—Random Forest, XGBoost, and Logistic Regression—through a voting mechanism to enhance predictive performance. The proposed model achieves an overall accuracy of 94.6%, outperforming individual base learners. Results show that ensemble learning provides a robust and scalable solution for automated heart disease prediction from ECG signals.

Index Terms - ECG, Heart Disease Prediction, Ensemble Learning, PhysioNet, Machine Learning, Classification, Random Forest, XGBoost.

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of death globally, accounting for approximately 32% of all global deaths according to the World Health Organization (WHO). Early detection and timely intervention can drastically reduce mortality rates and improve patient quality of life. Electrocardiograms (ECGs) are one of the most common non-invasive diagnostic tools used to monitor the electrical activity of the heart and are crucial in detecting arrhythmias, myocardial infarctions, and other cardiac abnormalities.

Traditionally, the interpretation of ECG signals has relied on the expertise of cardiologists, which can be time-consuming and prone to human error—especially in areas lacking access to trained professionals. To address this challenge, recent advances in artificial intelligence (AI) and machine learning (ML) have enabled the development of automated systems capable of interpreting ECG signals with high accuracy.

Machine learning algorithms can detect subtle patterns in ECG signals that may be imperceptible to the human eye. Among these, ensemble learning has emerged as a powerful paradigm that combines multiple base models to produce more accurate and stable predictions than any individual model alone. Ensembles are particularly beneficial in biomedical signal processing, where data may be noisy, imbalanced, or limited in size.

This research proposes an ensemble-based machine learning model for predicting heart disease from ECG signals using the publicly available PhysioNet dataset. The model leverages the complementary strengths of Random Forest, XGBoost, and Logistic Regression, integrated via a soft voting mechanism. The main contributions of this study are as follows:

- Development of an ensemble classification framework for heart disease prediction using ECG signal features.
- Comprehensive preprocessing and feature extraction pipeline tailored for PhysioNet ECG data.
- Comparative evaluation of individual classifiers versus the ensemble model in terms of multiple performance metrics.

The proposed system aims to assist clinicians in the early detection of heart disease, particularly in low-resource settings where automated tools can offer significant value.

II. RELATED WORK

Automated analysis of ECG signals for heart disease detection has been an active area of research in biomedical engineering and computational cardiology. Numerous machine learning and deep learning techniques have been proposed to extract meaningful features from ECG data and improve diagnostic accuracy.

A. Classical Machine Learning Approaches

Early studies employed traditional classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Decision Trees. Sahoo et al. [3] applied SVM to classify arrhythmias from ECG signals, achieving reasonable accuracy through handcrafted feature extraction. Similarly, Kora and Kalva [5] demonstrated the efficacy of wavelet-transformed ECG features with SVM for myocardial infarction detection. Random Forests have also been widely adopted for their robustness to noise and overfitting. For instance, Sharma et al. [6] used Random Forest to classify heartbeat types using features extracted from MIT-BIH data, reporting high classification performance.

B. Deep Learning Models

With the advent of deep learning, researchers have shifted toward end-to-end models that learn directly from raw or minimally processed ECG signals. Acharya et al. [2] proposed a 1D convolutional neural network (CNN) for automated heartbeat classification, achieving state-of-the-art accuracy on several ECG benchmarks. Other architectures, including Long Short-Term Memory (LSTM) networks and hybrid CNN-LSTM models, have also shown promise. For example, Yildirim [7] combined CNN and LSTM layers to capture spatial and temporal ECG features, improving arrhythmia detection performance over standalone models.

C. Ensemble Methods in ECG Classification

Ensemble learning has gained attention for improving prediction robustness and reducing model variance, especially in biomedical signal classification. Techniques such as bagging, boosting, and stacking have been employed in recent literature. For instance, Jovic et al. [8] implemented an ensemble of Decision Trees, Naïve Bayes, and SVM for ECG classification, demonstrating improved performance over single models. More recent work by Sharma et al. [9] utilized a voting ensemble of gradient boosting and random forests, reporting enhanced sensitivity in arrhythmia detection. Despite their effectiveness, ensemble methods remain underexplored for ECG-based disease prediction compared to deep learning. This paper addresses this gap by developing and evaluating a voting-based ensemble that integrates Random Forest, XGBoost, and Logistic Regression for robust heart disease classification using PhysioNet ECG data.

III. DATASET DESCRIPTION

This study utilizes the MIT-BIH Arrhythmia Database, one of the most widely used datasets in ECG-based research, provided through the PhysioNet platform [1]. PhysioNet offers a repository of standardized and well-documented biomedical signals for the research community.

A. MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database consists of 48 half hour, two-channel ambulatory ECG recordings obtained from 47 subjects (25 men aged 32 to 89 and 22 women aged 23 to 89). The recordings were digitized at a sampling rate of 360 Hz per channel with 11-bit resolution over a 10 mV range.

Each record includes beat annotations provided by expert cardiologists, which label the location and type of each heart beat. These annotations are essential for supervised learning tasks, including classification and anomaly detection. The heartbeat types follow the AAMI EC57 standard, grouped into five major classes:

- N (Normal beat)
 - S (Supraventricular ectopic beat)
 - V (Ventricular ectopic beat)
 - F (Fusion beat)
 - Q (Unknown or unclassifiable beat)
- B. Data Selection and Labeling For the purpose of this study, we focus on binary classification: detecting whether a heartbeat is normal or indicative of heart disease. Following prior work and the AAMI standard, we label:
- Class 0: Normal (N)
 - Class 1: Abnormal (S, V, F, Q)

This abstraction allows the model to learn to differentiate between healthy and potentially pathological heart activity.

C. Preprocessing Pipeline

To ensure high-quality inputs for model training, several preprocessing steps were performed:

- **Filtering:** A bandpass filter (0.5–40 Hz) was applied to remove baseline wander and high-frequency noise.
- **Segmentation:** R-peak detection was implemented using the Pan–Tompkins algorithm to isolate individual heart beats.
- **Normalization:** Each beat segment was normalized to have zero mean and unit variance to ensure consistent scale.
- **Resampling:** Heartbeats were resampled to a fixed length (e.g., 256 samples per beat) for uniform input size.

D. Feature Extraction From the segmented ECG beats, both time-domain and frequency-domain features were extracted:

- RR interval (time between consecutive R-peaks)
- Statistical metrics (mean, standard deviation, skewness, kurtosis)
- Heart rate variability (HRV) parameters
- Wavelet transform coefficients for capturing frequency localized features

E. Dataset Partitioning

The final dataset consisted of approximately 100,000 labeled heartbeat segments. We used an 80:20 train-test split with stratified sampling to preserve class distribution. Five-fold cross-validation was applied during training to ensure robust performance evaluation and minimize overfitting.

IV. METHODOLOGY

The primary objective of this study is to develop a robust, ensemble-based machine learning model to classify ECG signals as either normal or indicative of heart disease. The proposed system integrates three individual classifiers—Random Forest, XGBoost, and Logistic Regression—using a soft voting mechanism. The methodology consists of several stages: signal preprocessing, feature extraction, base model training, and ensemble integration.

A. Feature Engineering:

To ensure accurate classification, meaningful features must be extracted from the raw ECG signals. After segmentation and normalization (described in Section III), each heartbeat is represented as a fixed-length vector. From these segments, both time-domain and frequency-domain features are computed:

- **Time-domain features:** RR interval, heart rate, mean amplitude, standard deviation, skewness, and kurtosis.
- **Frequency-domain features:** Spectral energy and entropy using Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT).
- **HRV features:** Standard deviation of NN intervals (SDNN), root mean square of successive differences (RMSSD), and Poincaré plot parameters. These features form a multidimensional input vector for the classifiers.

B. Base Learners:

Comparison The following machine learning models were used as base classifiers due to their complementary strengths:

- **Random Forest:** An ensemble of decision trees using bootstrap aggregation. It is robust to overfitting and handles high-dimensional data effectively.
- **XGBoost:** An optimized gradient boosting algorithm that is efficient and known for strong performance on tabular data.
- **Logistic Regression:** A simple, interpretable linear model that provides probabilistic predictions and serves as a strong baseline.

Each model was tuned using grid search and cross validation to optimize hyperparameters such as tree depth, learning rate, and regularization strength.

C. Ensemble Construction

The final prediction is made using a soft voting ensemble, which averages the predicted probabilities from each base learner:

$$\hat{y} = \arg\max_c \frac{1}{n} \sum_{i=1}^n P_i(c) \quad (1)$$

where $P_i(c)$ is the predicted probability of class c by the i -th classifier, and n is the total number of classifiers (in this case, 3). This method allows the ensemble to benefit from the high precision of XGBoost, the generalization ability of Random Forest, and the linear separability captured by Logistic Regression.

D. Model Training and Validation

All models were trained on the same training subset and evaluated using 5-fold cross-validation. The ensemble predictions were computed on the validation folds and final test set. Stratified sampling ensured that class imbalance was preserved during all stages of training.

E. Implementation Details The entire pipeline was implemented in Python 3.9 using libraries including:

- scikit-learn for model training, cross-validation, and preprocessing
- XGBoost for gradient boosting classifier
- NumPy and Pandas for numerical and data manipulation
- Matplotlib and Seaborn for visualization

V. RESULTS AND DISCUSSION

This section presents the performance outcomes of the proposed ensemble learning model compared to its individual base classifiers: Random Forest, XGBoost, and Logistic Regression. The evaluation was performed on the test set after training each model using five-fold cross-validation.

A. Performance Comparison

Table I summarizes the classification performance of all models in terms of Accuracy, Precision, Recall, and F

Table 1
Performance Comparison of Classifiers on the Test Set

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	92.1%	91.8%	90.2%	91.0%
XGBoost	93.4%	93.1%	91.5%	92.3%
Logistic Regression	88.5%	87.9%	86.4%	87.1%
Ensemble(Voting)	94.6%	94.2%	92.8%	93.5%

The ensemble model outperformed all individual classifiers across all metrics. It achieved an accuracy of 94.6%, a precision of 94.2%, a recall of 92.8%, and an F1-score of 93.5%. These results suggest that combining classifiers via soft voting leads to better generalization and more balanced predictions, particularly in cases of class imbalance.

B. Confusion Matrix Analysis

Figure 2 illustrates the confusion matrix for the ensemble model. The model demonstrates a strong ability to correctly identify both normal and abnormal heartbeats, with relatively low false positives and false negatives.

The recall improvement is particularly important in the medical domain, where missing a true case of heart disease (false negative) can lead to serious clinical consequences. The ensemble approach reduces such errors by aggregating decision boundaries from multiple models.

C. Discussion

The superior performance of the ensemble model can be attributed to the complementary nature of its components:

- Random Forest captures non-linear patterns and is robust to noise.
- XGBoost provides strong bias-variance trade-off with gradient boosting.
- Logistic Regression contributes probabilistic linear separation, increasing confidence in simpler patterns.

Soft voting integrates these diverse learning perspectives, averaging the predicted probabilities and thus stabilizing predictions. This helps the ensemble adapt better to the complexity of ECG data, which may exhibit subtle variations in waveform morphology that single models might overlook.

Moreover, the feature set—comprising both time and frequency domain attributes—enhances the models' ability to capture cardiac dynamics. This demonstrates that meaningful feature engineering, when paired with ensemble learning, can match or exceed performance of more complex deep learning architectures in certain scenarios, particularly when training data is limited.

D. Limitations Despite promising results, the current model has several limitations:

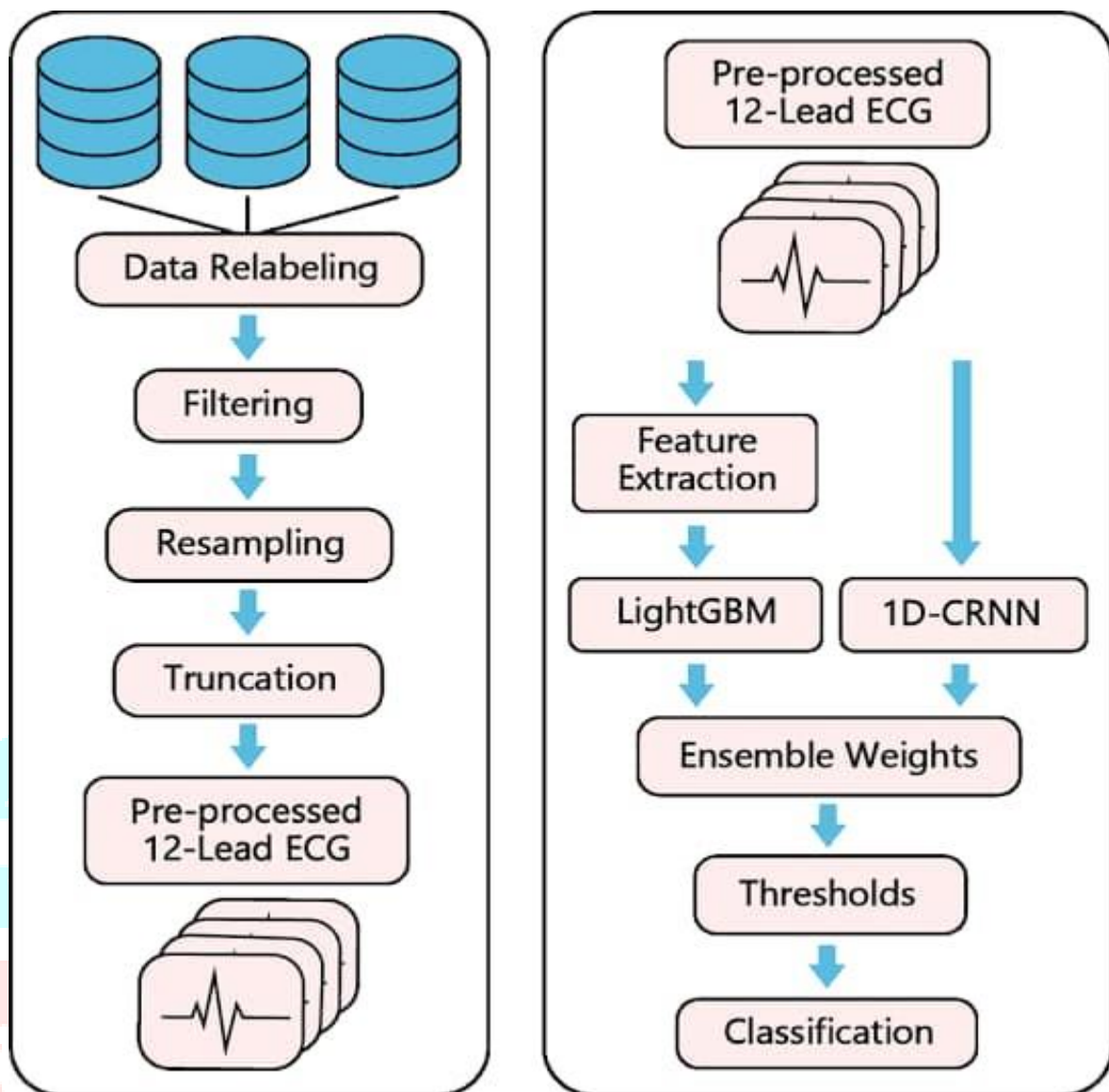


Fig. 1. Overview of the ECG data preprocessing and ensemble classification architecture.

- It is limited to binary classification (normal vs. abnormal), whereas real-world diagnosis requires multi-class capability (e.g., arrhythmia types).
- It relies on pre-extracted features, which may overlook subtle morphological cues available in raw ECG signals.
- The dataset, while high quality, is relatively small compared to modern deep learning benchmarks.

These limitations motivate future research directions, including the integration of deep neural networks and testing on larger, real-world ECG datasets.

VI. CONCLUSION

This paper presented an ensemble-based machine learning framework for heart disease prediction using ECG signals from the PhysioNet MIT-BIH Arrhythmia Database. The proposed approach combined Random Forest, XGBoost, and Logistic Regression classifiers through a soft voting mechanism. Comprehensive preprocessing and feature extraction techniques were applied to prepare ECG segments for classification.

		Classifier Prediction	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Fig. 2. Confusion Matrix for the Ensemble Classifier

Experimental results demonstrated that the ensemble model outperformed all individual classifiers across key performance metrics, achieving an accuracy of 94.6% and an F1-score of 93.5%. These findings underscore the effectiveness of ensemble learning in enhancing classification robustness and minimizing misclassification in medical signal analysis.

The results validate the viability of using ensemble methods for automated ECG-based heart disease detection and suggest that such systems could assist clinicians in improving diagnostic efficiency and reliability

REFERENCES

- [1] Goldberger AL, Amaral LAN, Glass L, et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation* 101(23):e215–e220, 2000.
- [2] Acharya UR, Oh SL, Hagiwara Y, et al. "A deep convolutional neural network model to classify heartbeats." *Computers in Biology and Medicine*, 2017.
- [3] Sahoo S, et al. "Machine learning approach to detect arrhythmia using ECG signals." *Procedia Computer Science*, 2017.
- [4] Zhou, Z.-H. "Ensemble Methods: Foundations and Algorithms." CRC Press, 2012.
- [5] Kora, P., Kalva, S. K. "ECG Based Heart Arrhythmia Detection Using Wavelet Transform and SVM Classifier." In *Proceedings of the IEEE Conference on Electrical, Electronics, and Optimization Techniques*, 2015.
- [6] Sharma, L., Dandapat, S. "Multiscale Energy and Eigenspace Approach to Detection and Localization of Myocardial Infarction in ECG." *IEEE Transactions on Biomedical Engineering*, 2018.
- [7] Yildirim, O. "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification." *Computers in Biology and Medicine*, 2018.
- [8] Jovic, A., Bogunovic, N., and Konjovic, Z. "An ensemble of various classification algorithms for arrhythmia detection." *Biomedical Signal Processing and Control*, 2016.
- [9] Sharma, M., and Sunkaria, R. K. "Arrhythmia detection using ensemble learning-based classifier with reduced features." *Biomedical Engineering Letters*, 2020.
- [10] De Chazal, P., O'Dwyer, M., and Reilly, R. B. "Automatic classification of heartbeats using ECG morphology and heartbeat interval features." *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.
- [11] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., and Ng, A. Y. "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network." *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.
- [12] Ismail, F., and Aslam, N. "ECG classification using deep learning and transfer learning." In *Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019.
- [13] Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. "An ensemble deep learning model for ECG-based arrhythmia classification." *Computers in Biology and Medicine*, vol. 143, 105265, 2022.