# ECG Arrhythmia Classification using Stacking model

Report 1
*Machine Learning in Medicine*

***Abstract*—This paper proposes a stacking ensemble learning framework for multi-class ECG heartbeat classification using the MIT-BIH Arrhythmia dataset. The framework combines heterogeneous base learners, including Extra Trees, Random Forest, Light Gradient Boosting Machine, and Logistic Regression, whose out-of-fold predictions are generated via 7-fold Stratified Cross-Validation and integrated by an XGBoost meta-model. Experimental results show that the proposed approach achieves an accuracy of 98.33% on the test set, with a macro-averaged precision of 94.03%, recall of 89.15%, and F1-score of 91.45%. Class-wise analysis using confusion matrix demonstrates robust performance across both majority and minority classes, confirming the effectiveness of the stacking strategy for ECG arrhythmia classification.**

## I. Introduction

Electrocardiogram (ECG) monitoring is widely used to detect cardiac arrhythmias. Although deep learning models such as Convolutional Neural Networks (CNNs) have shown strong performance in ECG classification, traditional machine learning models are still useful, especially for building baselines and working with structured data.

In this report, i use a stacking model to classify heartbeats into five categories. The ensemble combines several models, including Extra Trees (ET), Random Forest (RF), LightGBM (LGB), Logistic Regression (LGR), and XGBoost (XGB), to take advantage of their different strengths. This approach also helps deal with common challenges in ECG data, such as high dimensionality (187 features per sample) and severe class imbalance.

## II. Dataset Description

The dataset used in this study is based on the MIT-BIH Arrhythmia Database, a widely used benchmark for ECG classification tasks. The data has been pre-processed and segmented into individual heartbeats. Each heartbeat signal is normalized to reduce scale differences across samples.

### A. Data Structure

The dataset is provided in two CSV files:

- `mitbih_train.csv`: used for model training.
- `mitbih_test.csv`: used for model evaluation.

Each row corresponds to a single heartbeat and contains 188 columns. The first 187 columns represent ECG signal values sampled over time at a frequency of 125Hz. The final column contains the class label, represented as an integer from 0 to 4.

### B. Class Distribution and Imbalance

Normal heartbeats account for the majority of samples, while abnormal heartbeat types appear much less frequently. This imbalance makes the classification task more challenging, as standard training procedures tend to favor the majority class. The imbalance between classes is significant, with the ratio between the majority class (Normal) and the minority class (Fusion) being approximately 113:1. Without applying appropriate techniques to address this issue, a model could achieve around 82% accuracy by predicting only the Normal class, while failing to correctly identify abnormal heartbeats.

Table I shows the original class distribution in the training set.

TABLE I
Original Class Distribution (Training Set)

| ID | Code | Description | Count |
|---|---|---|---|
| 0 | N | Normal Beat | 72,471 |
| 1 | S | Supraventricular premature | 2,223 |
| 2 | V | Premature ventricular contraction | 5,788 |
| 3 | F | Fusion of ventricular & normal | 641 |
| 4 | Q | Unclassifiable beat | 6,431 |

## III. Implementation

This study implements a stacking ensemble learning framework for multi-class ECG heartbeat classification using the MIT-BIH Arrhythmia dataset. The entire experimental pipeline is developed in Python, utilizing the `scikit-learn`, `LightGBM`, and `XGBoost` libraries.

### A. Dataset Preparation

The MIT-BIH Arrhythmia dataset is provided in CSV format and consists of extracted ECG features with corresponding class labels. The dataset is divided into independent training and testing sets. For each sample, the final column represents the target class label, while the remaining columns correspond to input features.

### B. Base Learners

Four heterogeneous classifiers are employed as base learners in the stacking architecture:

- Extra Trees Classifier (ET)
- Random Forest Classifier (RF)
- Light Gradient Boosting Machine (LGBM)
- Logistic Regression (LGR)

The base learners include both linear and non-linear models, enabling diverse decision patterns and complementary error characteristics. All models are initialized with fixed random seeds to ensure reproducibility.

### C. Stacking Strategy

A Stratified K-Fold cross-validation strategy with $K = 7$ folds is employed to generate out-of-fold (OOF) predictions from each base learner. Stratification preserves the class distribution in each fold and ensures reliable meta-feature generation.

For each fold, the following steps are performed:

1) The base learners are trained on the training subset.
2) Class probability predictions are generated for the corresponding validation subset.
3) The predicted probabilities are stored as out-of-fold meta-features.

The OOF predictions from all base learners are concatenated to form the meta-feature matrix. This matrix is used exclusively for training the meta-model, thereby preventing information leakage.

### D. Meta-Model

XGBoost (XGB) is employed as the meta-model in the stacking framework. The meta-model learns how to optimally combine the probabilistic outputs of the base learners, capturing higher-order interactions between their predictions. By leveraging gradient boosting, the meta-model enhances overall classification performance and robustness.

After training on the OOF meta-features, the final stacking ensemble is evaluated on the independent test set.

### E. Evaluation Metrics

The performance of the stacking ensemble on the test set is evaluated using the following standard classification metrics:

- Accuracy
- Precision (macro-averaged)
- Recall (macro-averaged)
- F1-score (macro-averaged)

Macro-averaged metrics are reported to ensure balanced evaluation across both majority and minority classes.

## IV. Results and Discussion

This section presents the experimental results of the proposed stacking ensemble framework. Stratified 7-fold cross-validation is employed to generate out-of-fold predictions from the base models, which are later used to train the meta-classifier. The fold-wise cross-validation results are first reported to analyze the stability of individual models, followed by the evaluation of the final stacking ensemble on an independent test set.

TABLE II
Fold-wise Cross-Validation Accuracy of Models

| Fold | ET | RF | LGB | LGR |
|---|---|---|---|---|
| Fold 1 | 0.9779 | 0.9760 | 0.9799 | 0.9149 |
| Fold 2 | 0.9762 | 0.9739 | 0.9781 | 0.9128 |
| Fold 3 | 0.9762 | 0.9739 | 0.9791 | 0.9141 |
| Fold 4 | 0.9789 | 0.9779 | 0.9796 | 0.9142 |
| Fold 5 | 0.9766 | 0.9759 | 0.9791 | 0.9145 |
| Fold 6 | 0.9759 | 0.9743 | 0.9784 | 0.9140 |
| Fold 7 | 0.9754 | 0.9731 | 0.9775 | 0.9114 |
| **Mean** | 0.9767 | 0.9750 | **0.9788** | 0.9137 |
| **Std** | 0.0011 | 0.0015 | **0.0008** | 0.0011 |

### A. Cross-Validation Results of Base Models

Table II reports the classification accuracy obtained by each model across the 7 folds of Stratified Cross-Validation, together with the mean and standard deviation.

As shown in Table II, the tree-based models (ET, RF, and LGB) demonstrate consistently high and stable performance across all folds, with LightGBM achieving the highest mean accuracy and the lowest standard deviation. This indicates strong generalization ability and robustness to variations in the training data.

In contrast, Logistic Regression exhibits noticeably lower accuracy across all folds. However, its role in the proposed framework is not to act as a strong base classifier, but rather to serve as a meta-classifier that learns how to optimally combine the predictions of the base learners.

### B. Test Performance of the Stacking Ensemble

After training the Logistic Regression meta-classifier on out-of-fold predictions generated from the base models, the final stacking ensemble is evaluated on an independent test set. The results are summarized in Table III.

TABLE III
Test Performance of the Stacking Ensemble

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Stacking Model | 0.9833 | 0.9403 | 0.8915 | 0.9145 |

### C. Confusion Matrix and Class-wise Performance Analysis

TABLE IV
Class-wise Performance of the Stacking Ensemble

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Class 0 | 0.9868 | 0.9958 | 0.9913 | 18118 |
| Class 1 | 0.8933 | 0.7680 | 0.8259 | 556 |
| Class 2 | 0.9735 | 0.9385 | 0.9557 | 1448 |
| Class 3 | 0.8571 | 0.7778 | 0.8155 | 162 |
| Class 4 | 0.9905 | 0.9776 | 0.9840 | 1608 |
| Macro Avg | 0.9403 | 0.8915 | 0.9145 | 21892 |
| Weighted Avg | 0.9828 | 0.9833 | 0.9829 | 21892 |

Figure 1 illustrates the confusion matrix of the stacking ensemble on the test set. The majority of samples are correctly classified along the diagonal, indicating strong overall performance.

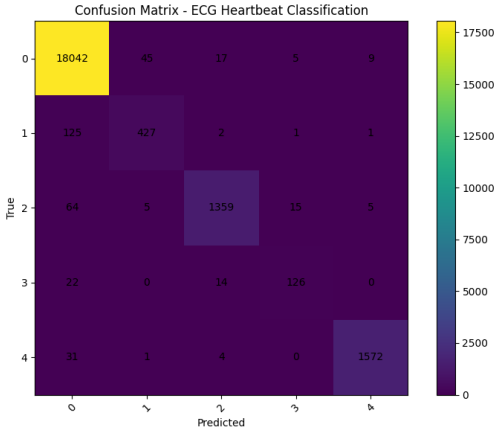


Fig. 1. Confusion matrix of the stacking ensemble on the test set.

Class 0 achieves the highest recall (99.58%), reflecting the model's ability to accurately recognize normal heartbeats. Classes 1 and 3 exhibit lower recall values, which can be attributed to class imbalance and overlapping feature distributions with other arrhythmia classes. Despite this challenge, the stacking ensemble maintains high precision across all classes, demonstrating robust decision boundaries.

These results confirm that the proposed stacking framework not only achieves high overall accuracy, but also maintains balanced performance across majority and minority classes.

### *D. Discussion*

The experimental results confirm that the stacking ensemble outperforms all individual models evaluated during cross-validation. Although Logistic Regression performs poorly as a standalone classifier, it effectively serves as a meta-learner by combining the complementary strengths of the non-linear base models.

The use of Stratified K-Fold cross-validation to generate out-of-fold predictions is essential for preventing information leakage and ensuring a reliable stacking procedure. Overall, the proposed ensemble framework achieves a strong balance between precision and recall, making it suitable for multi-class ECG arrhythmia classification and potential clinical applications.

## V. Conclusion

In this study, a stacking ensemble learning framework was proposed for multi-class ECG heartbeat classification. The framework integrates multiple tree-based base learners, including Extra Trees, Random Forest, and LightGBM, with Logistic Regression employed as a meta-classifier to combine their out-of-fold predictions.

Experimental results demonstrate that the proposed stacking model achieves strong and stable performance. The base models exhibit consistent accuracy across 7-fold Stratified Cross-Validation, indicating robust generalization. By leveraging the complementary strengths of the base learners, the stacking ensemble attains a test accuracy of 98.33%, with a macro-averaged F1-score of 91.45%, outperforming all individual models.

Further analysis using the confusion matrix and class-wise performance metrics confirms that the model effectively handles both majority and minority classes. While certain arrhythmia classes remain challenging due to class imbalance, the stacking approach improves overall robustness by reducing misclassification errors and maintaining high precision across all classes.

Overall, the proposed stacking ensemble provides an effective and reliable solution for ECG arrhythmia classification. Future work may explore advanced imbalance handling techniques, such as cost-sensitive learning or data augmentation, as well as the integration of deep learning-based feature extraction to further enhance classification performance.