MDPI

Article

Blending Ensemble Learning Model for 12-Lead ECGs-based Arrhythmia Classification

Hai-Long Nguyen [1], Van Su Pham [2] and Hai-Chau Le [1,3,*]

[1] Data and Intelligent Laboratory, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[2] Faculty of Electronics Engineering, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[3] Department of Data Engineering, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
* Correspondence: chaulh@ptit.edu.vn

Abstract: The increasing prevalence of heart diseases has driven the development of automated arrhythmia classification systems using machine learning and electrocardiograms (ECG). This paper presents a novel ensemble learning method for classifying multiple arrhythmia types using 12-lead ECG signals through a blending technique. The framework employs a predetermined meta-model from foundation models, while the remaining models serve as potential base estimators, ranked by accuracy. Using sequential forward selection and meta-feature augmentation, the system determines an optimal base estimator set and creates a meta-dataset for the meta-model, which is optimized through grid search with k-fold cross-validation. Experiments conducted with seven diverse machine learning algorithms (Adaptive boosting, Extreme gradient boosting, Decision trees, K-nearest neighbors, Logistic regression, Random Forest, and Support vector machine) demonstrate that the proposed blending solution, utilizing an LR meta-model with three optimal base models, achieves superior classification accuracy of 96.48%, offering an effective tool for clinical decision support.

Keywords: Machine learning; Ensemble learning; Blending; Electrocardiogram; Arrhythmia classification.

1. Introduction

Nowadays, cardiovascular diseases are among the most dangerous health issues, responsible for a significant number of deaths globally. The World Health Organization reports that cardiovascular diseases were responsible for 17.9 million deaths in 2019, representing 32% of all global deaths, with 85% of these cases stemming from heart attacks and strokes. [1]. While cardiovascular-related death rates have declined in developed countries due to improved healthcare and healthier lifestyles, they remain high in low- and middle-income countries [2]. The critical nature of these statistics underscores the urgent need for sophisticated diagnostic tools and early detection methods for cardiac irregularities. Among various diagnostic approaches, electrocardiogram (ECG) monitoring has established itself as the gold standard for arrhythmia detection, offering reliable insights into the heart's electrical patterns and potential abnormalities. This non-invasive technique provides valuable data for both immediate diagnosis and long-term cardiac monitoring. Thus, developing effective solutions for early identification of irregular heart rhythms is essential to reduce the impact of these diseases, and the electrocardiogram (ECG) is widely regarded as the most reliable method for detecting arrhythmias [3–5].

The evolution of cardiac diagnostics has been significantly enhanced by technological advancements, particularly in the realm of artificial intelligence. Recent developments have witnessed a surge in the application of sophisticated machine learning (ML) and deep learning methodologies for cardiac arrhythmia classification. Researchers have explored diverse algorithmic approaches, implementing classical machine learning techniques such as K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM)

Citation: Nguyen, H.-L.; Pham, V.-S; Le, H.-C. Blending Ensemble Learning for Enhanced Arrhythmia Classification Utilizing 12-Lead ECGs. *Computers* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

[6–8]. These methods have been deployed both independently and in innovative hybrid configurations, demonstrating promising results in detecting irregular heart rhythms [2,3,9]. The integration of these advanced computational techniques with traditional ECG analysis represents a significant step forward in improving both the accuracy and efficiency of arrhythmia detection, marking a new era in cardiovascular diagnostics.

Moreover, as the clinical standard, thanks to the advantages of offering a more comprehensive view of the heart's electrical activity by capturing signals from 12 different angles and enhancing the ability to diagnose complex arrhythmias, myocardial infarctions, and other cardiac conditions with greater accuracy and specificity, 12-lead ECGs are crucial for thorough cardiac assessments, allowing for more precise and reliable diagnoses than other simplified ECG configurations [5,10,11]. Recently, there have been many research attempts to classify arrhythmia based on 12-lead ECGs [12–14]. Authors in [15] proposed a convolutional neural network (CNN) method to transform ECG signals into RGB images using a continuous wavelet transform and then, features are extracted out of the images by applying CNN to generate a CNN feature representation in which the last layer is soft-max regression for multiclass classification. Especially, the work of [16] developed advanced 12-lead ECG signal processing techniques and two arrhythmia classification methods that are support vector machines and MLP. It's shown that the MLP outperforms the SVM in analyzing the diverse 12-lead ECG signals, achieving an accuracy of 84.2%. Another impressive work introduced in [17] demonstrated a high accuracy of 97.0% can be achieved for the classification of 4 arrhythmia classes. However, besides the ECG features, they both need further clinical features which may cause difficulties in applying to automated classification systems.

On the other hand, among recently advanced machine learning approaches, blending, an advanced ensemble learning technique, enhances predictive performance by combining the strengths of diverse models through a meta-model, which makes final detection decisions based on the multiple base models' outputs [18,19]. The meta-model is trained on a separate validation set to prevent overfitting and ensure generalizability. Popularized in competitions like the Netflix Prize and Kaggle, blending is valued for its versatility and effectiveness in complex predictive tasks. Although blending has been widely applied across various fields, there is currently no known application of this technique for 12-lead ECG arrhythmia classification.

In this paper, based on 12-lead ECG signals, we propose a novel ensemble learning approach for arrhythmia classification using 12-lead ECG signals through a blending strategy. The framework employs a predefined meta-model based on foundational architectures, while complementary models serve as candidate base estimators, ranked by their accuracy. The method optimizes performance through sequential forward selection and meta-feature enhancement, creating a refined meta-dataset for the meta-model, which is then fine-tuned using grid search with k-fold cross-validation. Experimental evaluation across seven machine learning algorithms—Adaptive Boosting, Extreme Gradient Boosting, Decision Trees, K-Nearest Neighbors, Logistic Regression, Random Forest, and Support Vector Machine—demonstrates that the proposed blending approach, combining a Logistic Regression meta-model with three optimized base models, achieves 96.48% classification accuracy, making it a robust tool for clinical decision support.

2. Data and Pre-processing

The Chapman ECG dataset, a comprehensive resource developed through a collaboration between Chapman University, Shaoxing People's Hospital, and Ningbo First Hospital, is considered the primary dataset for method development and validation of arrhythmia classification in this study [17]. The dataset includes 12-lead ECG recordings sampled at 500 Hz with a duration of 10 seconds from 10,646 patients. Expert clinicians have provided annotations for 11 common cardiac rhythms and 67 additional cardiovascular conditions. These features make the dataset a valuable resource for developing, comparing,

and optimizing statistical and machine-learning methods in the study of arrhythmias and cardiovascular disorders [16,20].

For automatic cardiac arrhythmia classification using 12-lead ECG signals, demographic features such as age and gender were not considered. Signal processing and feature extraction were performed on each ECG lead using the Neurokit2 Python library for neurophysiological signal processing [21]. ECG signals were cleaned to remove noise and improve peak-detection accuracy. In Neuroket, various cleaning methods are implemented, including *pantompkins1985*, *hamilton2002*, *elgendi2010*, *vg*, *biosppy*, and *neurokit*. Comprehensive experiments were conducted with different methods and the most suitable one, *neurokit*, was then selected. Herein, *neurokit* uses a 0.5 Hz high-pass Butterworth filter with the order of 5, followed by powerline filtering. After that, the feature extraction was performed utilizing the neurokit core functions such as peak detection, heart rate calculation, and QRS complex delineation. This process extracted 20 features, including R-peak count, mean, median, standard deviation, and range of RR intervals, skewness and kurtosis of RR intervals, mean R-peak amplitude compared to the isoelectric line, P-peak to R-peak ratio, median difference between R-R and P-P intervals, T-peak count, Q-peak count, mean and variance of QT, PR, PR-segment, and ST-segment intervals. Figure 1 illustrates these key ECG signal features. Features with excessive missing values were then eliminated, and the remaining missing data was imputed using probability density estimation.

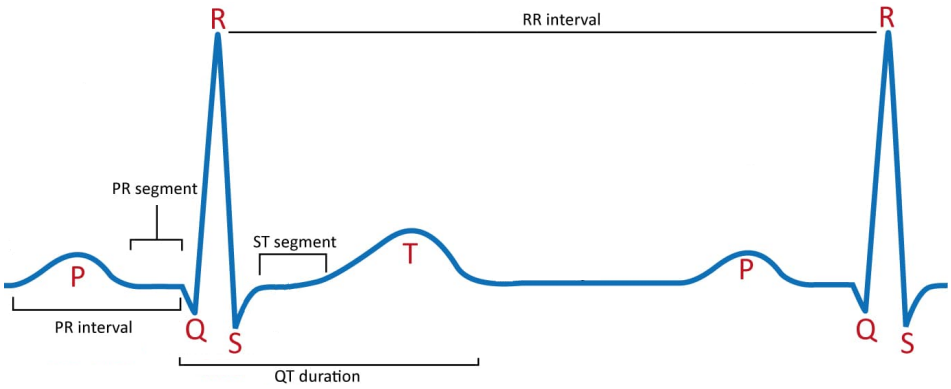

Figure 1. Typical ECG features considered

Based on their clinical significance, the 11 cardiac rhythms identified by expert clinicians were consolidated into four main categories: Atrial Fibrillation/Flutter (AFIB), Sinus Bradycardia (SB), Sinus Rhythm (SR), and General Supraventricular Tachycardia (GSVT) [12,16,17]. The AFIB category encompasses both atrial fibrillation and flutter conditions. SB represents cases of sinus bradycardia, while SR includes both regular and irregular sinus rhythms. The GSVT classification incorporates five distinct conditions: supraventricular tachycardia, atrial tachycardia, atrioventricular node reentrant tachycardia, atrioventricular reentrant tachycardia, and wandering atrial pacemaker. Besides reflecting shared treatment protocols and physiological mechanisms within each category, this consolidation also addresses several technical challenges in machine learning implementation: it mitigates class imbalance issues, enhances feature distinction between categories, improves model training efficiency, and promotes better generalization to real-world scenarios. Furthermore, the simplified classification system offers practical advantages such as reduced annotation complexity, higher inter-rater reliability among clinicians, more efficient real-time application, and decreased risk of misclassification between similar subtypes. This balanced approach ensures both clinical relevance and technical feasibility while maintaining the essential diagnostic utility of the classification system.

Finally, the attained values of the ECG features are normalized and scaled to a range [0, 1] utilizing the MinMaxScaler algorithm. The final dataset contained 10,646 samples, each characterized by 212 features. This dataset was divided into training and testing sets using

an 80:20 split ratio, with care taken to maintain consistent class distribution proportions across both subsets for the four arrhythmia categories.

3. Methodology

In this work, we develop an innovative ensemble learning method for classifying multiple types of arrhythmias using 12-lead ECG signals, as illustrated in Figure 2. The main idea of our approach lies in the strategic combination of diverse machine learning models through blending, a sophisticated ensemble technique that surpasses the capabilities of individual models [18]. The proposed method builds on the principles of ensemble learning but differs from conventional ensemble methods such as stacking, bagging, and boosting by employing a specialized meta-model architecture to synthesize predictions and combine the strengths of various base models. The implemented framework incorporates a set of the considered foundation machine learning models, denoted L, in which one meta-model, X, is predetermined and the others are considered as the candidates of base estimators, this set is named R. The algorithms in the set of base estimator candidates, R, are ranked in descending order based on the accuracy obtained on the given dataset. A sequential forward selection algorithm is applied to figure out the optimal base estimator set, $H = \{H_1, H_2, ..., H_m\}$ where m is the number of the selected base models. Each base model is then trained independently on the normalized ECG dataset, and their predictions are used to create a new dataset, called meta dataset, for the meta-model's input. The blender, a meta-model optimally chosen from the available algorithms, learns how to best combine predictions from the underlying models to create more accurate and stable classifications. The system uses grid search with k-fold cross-validation to find the most effective settings. In the final step, the meta-model makes the overall arrhythmia classifications based on the selected base models' outputs from the testing dataset.

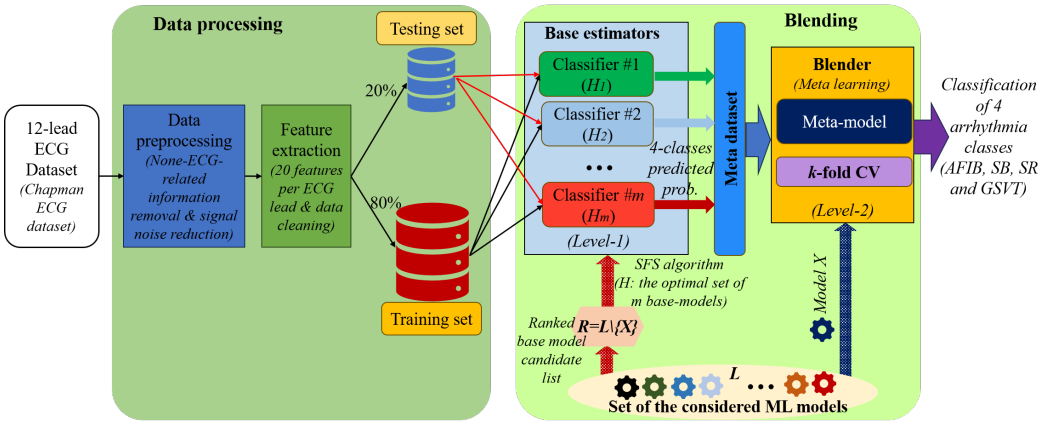

Figure 2. Flow of method development

The fundamental mechanism underlying our meta-learning aggregation procedure is meta-feature augmentation, wherein we supplement the validation dataset with predictor variables derived from the base models' outputs. The most important consideration pertains to the selection of an appropriate meta-model (level 2 ensemble aggregator) for feature synthesis. The ensemble architecture's efficacy is quantified through standardized performance metrics, with careful attention to preserving class distributions. This hierarchical integration methodology effectively addresses the inherent complexities of multi-class arrhythmia differentiation, yielding a diagnostic framework with enhanced discriminative capabilities and statistical reliability.

3.1. Base Model Selection and Training

To leverage the advantages of the blending ensemble learning approach, a predetermined set of machine learning algorithms that are popular, diverse, and efficient is considered. In fact, the base model selection criteria encompass historical performance

metrics, functional diversity, and domain-specific considerations. Each base model is independently trained on the training subset of the ECG dataset, which contains the normalized feature values and corresponding arrhythmia labels. This training aims to optimize each model's parameters to accurately classify four arrhythmia classes including Atrial Fibrillation (AFIB), Sinus Bradycardia (SB), Sinus Rhythm (SR), and Supraventricular Tachycardia (GSVT).

Following the training phase, all base models of the predetermined set generate predictive outputs on both Training and Testing cohorts, yielding either class-specific probability distributions or discrete class assignments, contingent upon meta-model specifications. The base models' Training set predictions are subsequently aggregated to construct a meta-feature space, effectively encoding the collective decision patterns into a higher-dimensional representation for meta-level learning.

3.2. Meta-Model Training and Validation

The meta-model, X which is actually selected as the remaining algorithm from the initial set of fundamental machine learning algorithms, is trained using the outputs of the base models as input features. This hierarchical architecture enables the meta-learner to discover optimal fusion strategies for the base predictions, thereby maximizing classification efficacy. The meta-learner's training corpus comprises the composite predictions generated by base estimators on the Training partition, paired with their corresponding ground truth labels.

The trained meta-learner subsequently processes the Testing partition by synthesizing the predictive outputs from all base estimators into final classification decisions. We implement k-fold cross-validation at the meta-level to ensure robust generalization of the fusion mechanisms. This architectural framework systematically leverages the complementary strengths of individual base estimators while mitigating their respective limitations, ultimately yielding enhanced discriminative performance across the four arrhythmia categories.

3.3. Development of Blending Models

Given the set of fundamental machine learning algorithms, L, we consider every blending configuration established by combining each model, X, of L as the meta-model ($X \in L$) and a set of base models, $R = L \setminus \{X\}$, denoted as $R - X$. Firstly, the set of base models is sorted in descending order based on the accuracy obtained for individual algorithms. Then, the searching algorithm, named *Best Base Model Combination* - which is based on a sequential forward selection strategy, is applied to determine the optimal blending configuration, $H - X$. Algorithm 1 presents a detailed formalization of the progressive ensemble construction methodology, outlining the systematic derivation of optimal base estimator compositions. The base model selection algorithm implements a sequential forward optimization strategy, wherein individual models are iteratively incorporated into an expanding active ensemble. This progressive architectural construction systematically evaluates candidate algorithms from the available pool, optimizing the ensemble composition through stepwise performance validation. Each iteration augments the active subset with the estimator (base model) that maximizes the ensemble's discriminative capacity, continuing until the predetermined size of R is achieved.

All possible blending configurations concerning the selection of every machine learning algorithm in L as the meta-model and the remaining algorithms as the base model candidates have been thoroughly investigated. The performance of the blending ensembles is evaluated using standard metrics including accuracy, precision, recall, F1-score, and specificity. The most efficient blending ensemble solution, in terms of accuracy, for the classification of the four cardiac arrhythmia classes is chosen based on the superior performance exhibited by the proper blending ensemble models.

Algorithm 1 Best Base Model Combination

Input:
- ECG dataset (training set and testing set)
- X: the meta-model
- $R = \{R_1, R_2, ..., R_T\}$: the set of considered machine learning algorithms where T is the element number of R

Output: $H = \{H_1, H_2, ..., H_m\}$: the optimal base model set where m is the number of models selected

(1) Ranking base models of the considered machine learning algorithm set, R, in descending order based on the accuracy over the dataset:
$R = \{R_1, R_2, ..., R_T \,|Acc_{R_i} \geq Acc_{R_j}, \forall\, 1 \leq i < j \leq T\}$ where Acc_* is the accuracy of $*$
(2) Finding the best base model combination:
$H = \emptyset$ ▷ Initial best set of selected base models
$Acc_{H-X} \leftarrow 0.0$ ▷ Initial accuracy of the best blending model where $H - X$ denotes the blending model combining the base model set of H and the meta-model of X
$H^\star = \emptyset$ ▷ The current considered combination of base models
for $i \leftarrow 1$ to T **do**
  $H^\star \leftarrow H^\star \cup \{R_i\}$
  Calculate $Acc_{H^\star - X}$
  if $Acc_{H^\star - X} > Acc_{H-X}$ **then**
    $Acc_{H-X} \leftarrow Acc_{H^\star - X}$
    $H \leftarrow H^\star$
  end if
end for
Return H

4. Experimental Results and Discussion

4.1. Experimental Setup and Performance Metrics

4.1.1. Experimental Setup

In this section, our blending ensemble framework strategically incorporates a diverse ensemble of fundamental machine learning algorithms, selected for their computational efficiency, algorithmic heterogeneity, and demonstrated predictive performance. To evaluate the effectiveness of the developed blending framework for 12-lead ECG-based arrhythmia classification involving four classes—Atrial Fibrillation, Sinus Bradycardia, Sinus Rhythm, and General Supraventricular Tachycardia, a set, L, of seven fundamental machine learning algorithms that are Adaptive Boosting (ADA), Decision Trees (DT), k-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), and XGBoost (XGB) ($L = \{ADA, DT, KNN, LR, RF, SVM, XGB\}$), is investigated. The proposed framework has been conducted on the considered basic ML pool, L in which six foundational models act as base predictors, combined with a meta-model. To leverage the advantages of the blending ensemble learning approach, seven basic machine learning algorithms that are popular, diverse, and efficient are considered. The process involves independent training of each base model on normalized ECG signals, followed by the aggregation of their predictions to construct a refined dataset for meta-model training. A concise overview of the algorithmic components selected for our ensemble framework is listed as follows.

1. AdaBoost (ADA) is an iterative ensemble learning algorithm for classification that trains base learners on weighted data, increasing the weights of misclassified points in each iteration, and makes final predictions based on a weighted combination of all learners, known for its high accuracy and versatility with various base learners.
2. Decision Tree (DT) is a supervised learning algorithm for classification and regression that splits data into branches based on features to form a tree-like model, offering interpretability but prone to overfitting on complex or noisy data.

3. k-Nearest Neighbors (KNN) is a non-parametric algorithm that predicts outcomes based on the labels or values of the k closest data points to a query point, using proximity in the feature space.
4. Logistic Regression (LR) is a supervised learning algorithm that uses the logistic (sigmoid) function to map inputs to probabilities for classification, relying on assumptions about the data distribution.
5. Random Forest (RF) is an extension of Bagging that builds multiple decision trees using random feature subsets to reduce overfitting, improve generalization, and handle high-dimensional data efficiently.
6. Support vector machine (SVM) is a supervised learning algorithm that classifies data by finding the hyperplane with the maximum margin between classes, using kernel functions to handle non-linearly separable data, thus enhancing generalization performance.
7. XGBoost (XGB) is a gradient boosting algorithm optimized for speed and performance, known for high accuracy and advanced regularization to prevent overfitting, with parallel processing capabilities for large datasets.

We comprehensively evaluated all potential blending configurations by systematically exploring meta-model selections from set L. With six base classifiers along with the output of four probabilities (four classes), the meta dataset includes 24 features and feeds to the meta-model. We employ k-fold cross-validation with $k = 3$, strategically selected to optimize performance validation and mitigate potential overfitting. Performance assessment utilized standard classification metrics, with the optimal ensemble configuration being selected based on superior classification performance across the four arrhythmia categories.

4.1.2. Performance Metrics

The five most important performance metrics including accuracy, F1-score, precision, recall, and specificity are utilized to evaluate the performance of our proposed blending framework for the classification of four arrhythmia types. Accuracy measures the overall proportion of correctly classified instances, providing a general sense of the model's performance across all classes. However, when class distribution is imbalanced, the F1-score becomes crucial as it balances precision and recall, ensuring that both false positives and false negatives are considered, especially for rare arrhythmias. In addition, precision indicates the proportion of correctly identified positive cases within all predicted positives, helping assess the model's reliability in detecting specific arrhythmias while recall measures the proportion of actual positives correctly identified, ensuring the model captures most of the pathological cases, which is critical in medical diagnostics. On the other hand, specificity assesses how well the model identifies negative cases, minimizing the risk of false alarms for non-pathological rhythms. These metrics are selected to provide a balanced evaluation of the model's performance, ensuring accurate detection, minimal false positives, and adequate sensitivity to arrhythmic conditions, which is essential for reliable ECG-based diagnosis of multi-class arrhythmias.

In a multi-class classifier, the performance metrics including Accuracy, Precision, Recall, F1-score, and Specificity are computed based on confusion matrix values. Denote the major components for each arrhythmia class, i ($i \in [AFIB, SB, SR, GSVT]$), of a multi-class confusion matrix respectively as True Positives (TP_i), False Positives (FP_i), False Negatives (FN_i), and True Negatives (TN_i). Here, TP_i represents correct predictions for class i, FP_i denotes incorrect predictions as class i, FN_i indicates instances of class i predicted as others, and TN_i includes all other correct predictions outside the class of interest. Formulas of the accuracy, Precision, Recall, F1-score, and Specificity can be expressed respectively as follows.

$$\text{Accuracy}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i + \text{TN}_i} \tag{1}$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{2}$$

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3}$$

$$\text{F1-Score}_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{4}$$

$$\text{Specificity}_i = \frac{\text{TN}_i}{\text{TN}_i + \text{FP}_i} \tag{5}$$

On the other hand, in machine learning, performance metrics such as F1-score, precision, recall, and specificity are often aggregated across multiple classes using macro, micro, and weighted averages to provide a comprehensive view of model performance. Macro average calculates the metric independently for each class and then takes the arithmetic mean, treating all classes equally regardless of their size, making it useful when class distribution is balanced. Micro average aggregates the contributions of all classes by summing up true positives, false positives, and false negatives across classes before calculating the metric, emphasizing the performance on the most frequent classes. This is particularly useful for imbalanced datasets where larger classes dominate the predictions. Weighted average computes the metric for each class, similar to the macro average, but weights each class's score by its support (the number of instances in that class), balancing the influence of both large and small classes. These averaging methods help capture different perspectives of model performance, ensuring more informed evaluation in multi-class or imbalanced classification scenarios.

4.2. Basic Model Training and Ranking

All seven fundamental machine learning algorithms in the given list, L, are considered as possible base models in blending ensemble models for 12-lead ECG-based classification of 4 arrhythmia types including AFIB, SB, SR, and GSVT. These base models are trained independently on the training data to generate the predicted probabilities, which are then used as input features for a meta-model, of the four arrhythmia classes, and the final performances on the testing set are evaluated. The algorithms are then ranked in descending order based on the attained accuracy score, as summarized in Table 1. It is important to note that hyper-parameter tuning is performed using Grid Search to identify the optimal model and its best configuration. The results show that RF achieved the highest accuracy at 96.20%, followed closely by XGB at 96.15%. KNN and SVM both achieved above 94+% accuracy while DT, ADA, and LR performed relatively lower but still achieved respectable accuracies above 91+%, with LR having the lowest accuracy at 91.88%.

Table 1. Ranked list of basic machine learning algorithms

The order	Machine learning algorithm	The obtained accuracy
1	RF	0.9620
2	XGB	0.9615
3	KNN	0.9526
4	SVM	0.9498
5	DT	0.9286
6	ADA	0.9225
7	LR	0.9188

The role of the basic ML models is to capture diverse patterns and strengths from the data, and their combined outputs are utilized by the meta-model to learn how to optimally integrate these predictions, enhancing overall blending model performance. Based on this ranked list, excluding the pre-selected meta-model, six other algorithms are utilized as the

candidates of base models to build blending configurations for determining the optimal blending model with the pre-determined meta-model.

4.3. Blending Model Development

From the ranked basic machine learning algorithms, blending ensembles are constructed by combining a meta-model, X, with the optimal base model set, H, selected using the *Best Base Model Combination* on the list, R, of 6 remaining fundamental algorithms in L. Consequently, seven blending ensembles are created with meta-models of ADA, DT, KNN, LR, RF, SVM, and XGB, respectively. For each blending ensemble, a meta-dataset is created by merging the predicted probabilities from the optimal set of classifiers that serve as base models at level 1. The meta-model is then trained on this meta-dataset along with the true labels from the training data. Its objective is to learn the optimal combination of base models' predictions to enhance overall performance. After training, the meta-model utilizes the base models' predictions on the testing set to make final predictions. The meta-model's performance is then evaluated using appropriate metrics, ensuring effective integration of the diverse outputs from base models to achieve improved accuracy, generalization, and robustness in predictions.

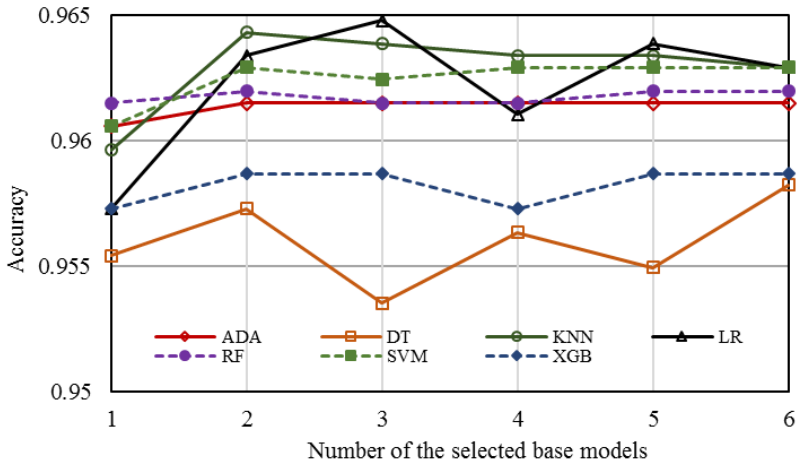

Figure 3. Based model set selection

The performance comparisons of the blending ensembles with different meta-models of ADA, DT, KNN, LR, RF, SVM, and XGB, for arrhythmia classification across four classes: AFIB, SB, SR, and GSVT, in terms of accuracy, with respect to the size of base model set is demonstrated in Figure 3. Most algorithms show relatively consistent performance, with accuracies ranging between 95.35% and 96.43%. KNN achieved the highest accuracy of 96.43% with only 2 base models, while DT showed the lowest best accuracy of 95.35% with 3 base models. Generally, increasing the number of base models didn't consistently improve accuracy across all algorithms, suggesting that model complexity doesn't necessarily correlate with better performance in this case. This implies that the performance of blending ensembles with the pre-determined meta-model strongly relies on the selected base model set. The optimal base model set sizes of the meta-models of ADA, DT, KNN, LR, RF, SVM, and XGB are achieved at 2 (96.15%), 6 (95.82%), 2 (96.43%), 3 (96.48%), 5 (96.20%), 4 (96.29%), and 2 (95.87%) respectively. Obviously, the blending configuration with the meta-model of LR offers the best performance in terms of accuracy.

In addition, Figure 4 shows the F1-score comparison of the optimal blending configuration with different meta-models. The graphs of three averaging F1-score metrics (macro, micro, and weighted), of seven blending ensembles with the meta-models of ADA, DT, KNN, LR, RF, SVM, and XGB demonstrate that the blending ensemble with LR meta-model consistently performs best across all metrics with scores around 96.4%, while that of XGB shows relatively lower performance with scores around 95.3-95.8%. The micro and

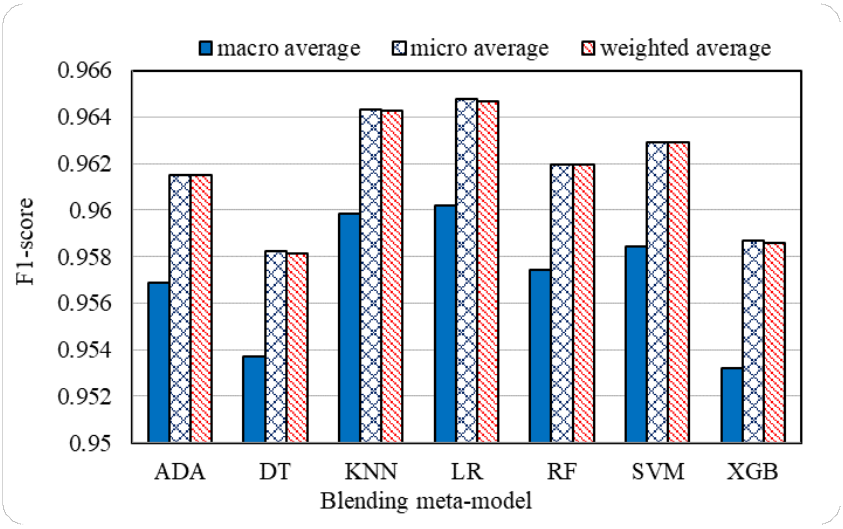

Figure 4. F1-score comparison of the optimal blending configuration with different meta-models

weighted averages generally show higher scores than macro averages across all models, suggesting better performance on more frequently occurring classes.

Table 2. Accuracy Comparison

Comparative configuration		Classification accuracy					Number of base models
		Overall	*Each arrhythmia class*				
			AFIB	***SB***	***SR***	***GSVT***	
Single machine learning algorithm	*ADA*	0.9225	0.9305	0.9850	0.9779	0.9516	-
	DT	0.9286	0.9404	0.9869	0.9704	0.9596	-
	KNN	0.9526	0.9690	0.9864	0.9779	0.9718	-
	LR	0.9188	0.9502	0.9756	0.9634	0.9484	-
	RF	0.9620	0.9704	0.9915	0.9864	0.9756	-
	SVM	0.9498	0.9676	0.9873	0.9793	0.9653	-
	XGB	0.9615	0.9690	0.9906	0.9892	0.9742	-
Blending framework with the pre-determined meta model	***ADA***	*0.9615*	*0.9690*	*0.9906*	*0.9892*	*0.9742*	*2*
	DT	*0.9582*	*0.9695*	*0.9873*	*0.9864*	*0.9732*	*6*
	KNN	*0.9543*	*0.9714*	*0.9920*	*0.9897*	*0.9756*	*2*
	LR	*0.9648*	*0.9723*	*0.9925*	*0.9892*	*0.9756*	*3*
	RF	*0.9620*	*0.9965*	*0.9906*	*0.9892*	*0.9746*	*5*
	SVM	*0.9629*	*0.9704*	*0.9911*	*0.9892*	*0.9751*	*4*
	XGB	*0.9587*	*0.9676*	*0.9915*	*0.9840*	*0.9742*	*2*

Moreover, Table 2 summarizes a detailed performance comparison of two approaches for arrhythmia classification: single machine learning algorithms and blending frameworks. In the single algorithm category, RF performed best with 96.20% overall accuracy, followed closely by XGB (96.15%). When using the blending framework, LR as a meta-model achieved the highest overall accuracy (96.48%) using 3 base models, with improved performance across all arrhythmia classes (AFIB, SB, SR, and GSVT). Both approaches show particularly significant performance in classifying Sinus Bradycardia and Sinus Rhythm, with accuracies consistently above 97+%. This means that our proposed blending framework outperforms traditional approaches and attains the highest performance with the meta-model of LR and the optimal set of only three base models. The blending design consistently achieves the highest performance, with an overall score of 96.48% and the highest accuracy in most individual classes, particularly SR (98.92%), SB (99.25%), and GSVT (97.56%). The performance across all meta-models is relatively similar, with only slight variations in accuracy, indicating that each algorithm captures the predictive patterns effectively. However, the blending models with the meta-model of DT and XGB generally show slightly lower overall scores, suggesting they may be less optimal choices compared to others like those of LR or KNN. These results highlight the blending ensemble

with the LR meta-model and the optimal set of three base models as a robust arrhythmia classification, providing balanced and high performance across all arrhythmia classes.

4.4. Proposed Blending Model

Based on the above observations, the blending design, which is proposed for the classification of four arrhythmia types in this work, contains an LR meta-model and a subset of 3 base models selected from the given seven fundamental machine learning algorithms. Here, the LR meta-model combined with CV procedure is implemented with different k values ranging from 3 to 10. The value of k for which the blending model produces the highest classification accuracy is selected as the optimal parameter. The performance of the proposed blending ensemble with the LR meta-model is given in Table 3. The proposed blending model demonstrates excellent performance in classifying four types of arrhythmia, with SB achieving the highest accuracy of 99.25% and best F1-score of 99.49%, followed by SR with 98.92% accuracy and 97.53% F1-score. The classes of AFIB and GSVT also show strong performance with accuracies of 97.23% and 97.56% respectively, and F1-scores above 93+%. The model's overall performance is robust, as evidenced by consistently high macro (96.02%), micro (96.48%), and weighted (96.47%) averages across all metrics, including precision, recall, and specificity, indicating reliable and balanced classification capabilities across all arrhythmia types. Additionally, the ROC-AUC curves of individual classes are also illustrated in Figure 5. Our developed model attains a significant AUC for all classes, where the highest AUC of 99.899% is achieved for the SB classification and the lowest AUC of 99.142% is for that of AFIB.

Table 3. Proposed blending model performance

Classification		**Parameters**				
		Accuracy	***F1-score***	***Precision***	***Recall***	***Specificity***
Individual class	*AFIB*	0.9723	0.9303	0.9367	0.9335	0.9834
	SB	0.9925	0.9949	0.9847	0.9898	0.9911
	SR	0.9892	0.9753	0.9731	0.9742	0.9929
	GSVT	0.9756	0.9372	0.9496	0.9434	0.9862
Overall	*Macro ave.*	0.9648	0.9602	0.9610	0.9594	0.9882
	Micro ave.		0.9648	0.9648	0.9648	0.9881
	Weighted ave.		0.9647	0.9646	0.9648	0.9887

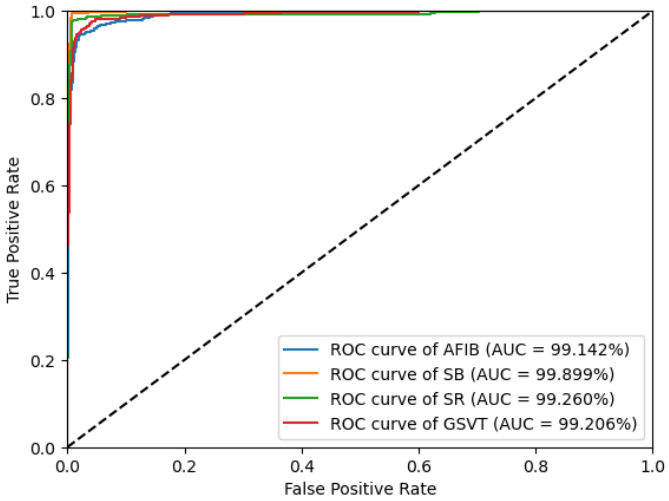

Figure 5. ROC-AUC score of the proposed blending model over the given seven ML algorithm set

4.5. Discussion

Our proposed blending framework can classify multiple arrhythmia types using 12-lead ECG signals efficiently by exploiting blending incorporating a sequential forward

selection of base models. The use of the sequential forward selection strategy helps minimize the number of base models and eliminating unnecessary ones while ensuring the highest performance. This results in a simpler algorithm complexity and makes the proposal more feasible to apply in practical environments. The blending framework also employs meta-feature augmentation with a meta-dataset created based on the outputs' of the optimal base models. By performing comprehensive experiments over the given fundamental machine learning algorithms, all possible blending configurations have been investigated to figure out the most efficient solution.

Over the set of the given seven basic machine learning algorithms, the best blending ensemble solution has been found to have the LR meta-model combining with the optimal set of only three base models for multi-classes arrhythmia classification using 12-lead ECGs. Without the necessity of the clinical features, i.e. age or gender, ..., our proposed solution significantly outperforms the notable conventional works given in [16] under the same conditions. As shown in Table 4, the proposed method achieves the highest overall accuracy of 96.48%, dramatically surpassing that of [16] (84.20%), representing a 14.58% improvement in terms of the overall accuracy. The framework demonstrates consistently high performance across all metrics—F1-scores, Precision, and Recall—indicating balanced and robust classification capabilities. The reason behind the success of our method comes from the efficiency of the blending which is a powerful ensemble technique that can significantly boost the performance of machine learning models by leveraging the complementary strengths of diverse algorithms. It is important to mention that the way we set up the cross-validation procedure also contributes significantly to our final results.

Table 4. Performance comparison

Performance parameters		***Conventional work [16]***	***The proposed method***	***Gaining percentage***
Overall accuracy		0.842	0.9648	14.58%
F1-score	*AFIB*	0.830	0.9303	12.09%
	SB	0.889	0.9949	11.91%
	SR	0.964	0.9753	1.17%
	GSVT	0.897	0.9372	4.48%
Precision	*AFIB*	0.883	0.9367	6.08%
	SB	0.879	0.9847	12.03%
	SR	0.941	0.9731	3.41%
	GSVT	0.895	0.9496	6.10%
Recall	*AFIB*	0.782	0.9335	19.37%
	SB	0.898	0.9898	10.22%
	SR	0.988	0.9742	-1.40%
	GSVT	0.899	0.9434	4.93%

5. Conclusions

Blending models represent a powerful yet accessible approach to ensemble learning, combining the predictive strengths of multiple models to achieve superior performance. In this paper, we propose an efficient ensemble learning method for classifying multiple arrhythmia types using 12-lead ECG signals through a blending technique that combines diverse machine learning models. The framework employs a predetermined meta-model from a set of foundation models, while the remaining models serve as potential base estimators, ranked by their accuracy. Using sequential forward selection, an optimal base estimator set is determined, and these base models are trained independently on normalized ECG data to create a meta-dataset. The system employs meta-feature augmentation and is optimized through grid search with k-fold cross-validation to learn the best combination of base model predictions. Extensive experiments were conducted using seven fundamental machine learning algorithms (ADA, DT, KNN, RF, SVM, XGB, and LR), evaluating all possible blending configurations. The most efficient solution, utilizing an LR meta-model with three optimal base models, achieved a superior classification accuracy of 96.48%. This approach not only improves interpretability and generalization but also reduces overfitting

risk, demonstrating its potential as a reliable tool for automated arrhythmia classification in clinical decision support.

Author Contributions: Conceptualization, H.-C.L. and V.-S.P.; methodology, H.-L.N. and H.-C.L.; formal analysis, H.-L.N.; investigation, H.-C.L. and V.-S.P.; writing—original draft preparation, H.-L.N.; writing—review and editing, H.-C.L. and V.-S.P.; supervision, H.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The dataset generated and/or analyzed during the current study is available on physionet.org. (https://physionet.org/content/ecg-arrhythmia/1.0.0/).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Organization, W.H. Cardiovascular Diseases. https://www.who.int/europe/news-room/fact-sheets/item/cardiovascular-diseases, 2024. Accessed: 15 August 2024.
2. Liu, J.; Li, Z.; Jin, Y.; Liu, Y.; Liu, C.; Zhao, L.; Chen, X. A review of arrhythmia detection based on electrocardiogram with artificial intelligence. *Expert review of medical devices* **2022**, *19*, 549–560.
3. Sahoo, S.; Dash, M.; Behera, S.; Sabut, S. Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey. *Irbm* **2020**, *41*, 185–194.
4. Merdjanovska, E.; Rashkovska, A. A framework for comparative study of databases and computational methods for arrhythmia detection from single-lead ECG. *Scientific Reports* **2023**, *13*, 11682.
5. Macfarlane, P.W.; Van Oosterom, A.; Pahlm, O.; Kligfield, P.; Janse, M.; Camm, J. *Comprehensive electrocardiology*; Springer Science & Business Media, 2010.
6. Li, J.; Pang, S.p.; Xu, F.; Ji, P.; Zhou, S.; Shu, M. Two-dimensional ECG-based cardiac arrhythmia classification using DSE-ResNet. *Scientific Reports* **2022**, *12*, 14485.
7. Jin, Y.; Li, Z.; Wang, M.; Liu, J.; Tian, Y.; Liu, Y.; Wei, X.; Zhao, L.; Liu, C. Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis. *Communications Medicine* **2024**, *4*, 31.
8. Qananwah, Q.; Ababneh, M.; Dagamseh, A. Cardiac arrhythmias classification using photoplethysmography database. *Scientific Reports* **2024**, *14*, 3355.
9. Dinakarrao, S.M.P.; Jantsch, A.; Shafique, M. Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques. *ACM Computing Surveys (CSUR)* **2019**, *52*, 1–37.
10. Jeong, D.U.; Lim, K.M. Convolutional neural network for classification of eight types of arrhythmia using 2D time-frequency feature map from standard 12-lead electrocardiogram. *Scientific reports* **2021**, *11*, 20396.
11. Andayeshgar, B.; Abdali-Mohammadi, F.; Sepahvand, M.; Almasi, A.; Salari, N. Arrhythmia detection by the graph convolution network and a proposed structure for communication between cardiac leads. *BMC Medical Research Methodology* **2024**, *24*. https://doi.org/10.1186/s12874-024-02223-4.
12. Jianwei Zheng, Huimin Chu, D.S.; et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports* **2020**, *10*, 2898.
13. Hajianfar, G.; Khorgami, M.; Rezaei, Y.; Amini, M.; Samiei, N.; Tabib, A.; Borji, B.K.; Kalayinia, S.; Shiri, I.; Hosseini, S.; et al. Comparison of Machine Learning Algorithms Using Manual/Automated Features on 12-Lead Signal Electrocardiogram Classification: A Large Cohort Study on Students Aged Between 6 to 18 Years Old. *Cardiovascular Engineering and Technology* **2023**, *14*, 786–800.
14. Yang, X.; Ji, Z. Automatic Classification Method of Arrhythmias Based on 12-Lead Electrocardiogram. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23094372.
15. Al Rahhal, M.M.; Bazi, Y.; Al Zuair, M.; Othman, E.; BenJdira, B. Convolutional neural networks for electrocardiogram classification. *Journal of Medical and Biological Engineering* **2018**, *38*, 1014–1025.
16. Aziz, S.; Ahmed, S.; Alouini, M.S. ECG-based machine-learning algorithms for heartbeat classification. *Scientific reports* **2021**, *11*, 18738.
17. Zheng, J.; Zhang, J.; Danioko, S.; Yao, H.; Guo, H.; Rakovski, C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data* **2020**, *7*, 48.
18. Gautam, K. *Ensemble methods for machine learning*; Manning, 2023.
19. Yao, J.; Zhang, X.; Luo, W.; Liu, C.; Ren, L. Applications of Stacking/Blending ensemble learning approaches for evaluating flash flood susceptibility. *International Journal of Applied Earth Observation and Geoinformation* **2022**, *112*, 102932.
20. Cao, Y.; Geddes, T.A.; Yang, J.Y.H.; Yang, P. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence* **2020**, *2*, 500–508.
21. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S.A. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* **2021**, *53*, 1689–1696. https://doi.org/10.3758/s13428-020-01516-y.