Array 32 (2026) 101366

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

MOE-ECG: Multi-objective ensemble fusion for robust atrial fibrillation detection using electrocardiograms

Abdolrahman Peimankar [a,*], Naser Hossein Motlagh [b], Smith K. Khare [c], Nicolai Spicher [d], Helena Dominguez [e], Vahid Abolghasemi [f], Koichi Fujiwara [g,h], Daniel Teichmann [a], Rahim Rahmani [i], Sadasivan Puthusserypady [d]

[a] SDU Health Informatics and Technology, The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, 5230, Denmark
[b] Department of Computer Science, University of Helsinki, Helsinki, 00100, Finland
[c] SDU Applied AI and Data Science, The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, 5230, Denmark
[d] Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark
[e] Bisperbjerg-Frederiksberg Hospital, Copenhagen, 2400, Denmark
[f] School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, UK
[g] Research Institute for Electronic Science, Hokkaido University, N20W10, Kita-Ward, Sapporo, 001-0020, Hokkaido, Japan
[h] Medilux Research Center, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, 630-0192, Nara, Japan
[i] Department of Computer and Systems Sciences, Stockholm University, Kista, 164 07, Sweden

ARTICLE INFO

Dataset link: MIT-BIH Atrial Fibrillation Database, Saitama Heart Database Atrial Fibrillation, Long-Term Atrial Fibrillation Database

Keywords:
Atrial fibrillation
Electrocardiogram (ECG)
Ensemble learning
Multi-objective optimization
Classifier diversity

ABSTRACT

Atrial fibrillation (AFib) is the most common sustained cardiac arrhythmia and imposes a major clinical and economic burden worldwide. Early and reliable detection is essential for reducing adverse outcomes, particularly as ECG data are increasingly generated by clinical and wearable monitoring devices. However, manual inspection of large ECG volumes is time-consuming and may delay timely decision-making. This paper proposes MOE-ECG, a multi-objective ensemble selection and fusion framework for robust AFib detection. The framework uses multi-objective particle swarm optimization to construct ensembles that jointly optimize predictive performance and inter-model diversity. RR intervals were derived from reference beat annotations, segmented into 20-, 60-, and 120-heartbeat windows with 50% overlap, and represented using statistical and nonlinear RRI features. From 15 extracted features, ten were retained using a two-stage relevance–redundancy correlation-based procedure. The framework used the MIT-BIH Atrial Fibrillation Database (AFDB) for training, the Saitama Heart Database Atrial Fibrillation (SHDB-AF) for validation, and the Long-Term Atrial Fibrillation Database (LTAFDB) for independent testing, with five runs used to assess stability. The 60-beat configuration provided the most balanced performance, achieving mean accuracy of 84.7%, precision of 69.7%, sensitivity of 98.2%, F_1-score of 81.5%, and AUC of approximately 0.95 on LTAFDB. MOE-ECG also showed higher mean accuracy, precision, and F_1-score than the single-best classifier and unoptimized average ensemble. These findings demonstrate competitive cross-dataset AFib detection with particularly high sensitivity on the independent test set, while retaining a compact and interpretable RRI-based representation across heterogeneous recording conditions.

1. Introduction

Atrial fibrillation (AFib) is the most common sustained cardiac arrhythmia and represents a major global health burden, which affects over 33 million people worldwide [1]. The prevalence increases with age and comorbidities such as hypertension, diabetes, and heart failure, and it continues to be one of the leading causes of stroke and heart-related morbidity and mortality worldwide [2,3]. Its intermittent and often asymptomatic nature poses a significant challenge to diagnosis, since many patients remain undiagnosed until severe complications occur. Therefore, early and reliable detection of AFib can reduce adverse clinical outcomes and substantially alleviate healthcare costs.

Traditionally, AFib diagnosis has relied on visual inspection of electrocardiograms (ECG) by clinicians to identify patterns such as irregular RR-intervals and absence of P waves [4,5]. However, the generation of large amounts of ECG data due to continuous monitoring,

* Corresponding author.
E-mail address: abpe@mmmi.sdu.dk (A. Peimankar).

https://doi.org/10.1016/j.array.2026.101366
Received 27 July 2026; Received in revised form 30 September 2026; Accepted 14 October 2026
Available online 22 October 2026
2590-0056/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

which is needed for the detection of paroxysmal AFib, makes a manual review impractical [6]. Next to the traditional 10-s 12-lead ECG measurement, Holter monitors and cardiac patches that continuously record ECG signals over multiple days, as well as handheld devices and smartwatches, have entered clinical practice [7]. These developments have increased the need for computationally efficient ECG processing and analysis methods suitable for wearable and edge devices [8]. This has motivated the development of automated and intelligent systems capable of accurate and efficient AFib detection from ECG recordings, including lightweight architectures designed for energy-constrained wearable devices [9,10]. Nevertheless, ECG recordings acquired during continuous and ambulatory monitoring remain susceptible to baseline wander, motion artifacts, muscle activity, and power-line interference, making reliable signal preprocessing an important component of automated analysis [11]. Over the past two decades, different signal processing and machine learning (ML) techniques have been proposed to assist in automatically detecting AFib using time-, frequency-, and nonlinear-domain features of RR-intervals or raw ECG signals [8,11–14].

While common ML algorithms such as support vector machines, random forests, and boosting ensembles have shown promising results when combined with carefully engineered features [15–17], most existing approaches either rely on single classifiers or employ fixed or homogeneous ensembles optimized for a single objective (typically classification accuracy). More recently, optimization-guided feature selection and heterogeneous ensemble fusion have shown that selecting complementary models and features can outperform the indiscriminate combination of all available learners in medical prediction tasks [18]. In many cases, ensemble construction is heuristic or based on simple averaging, without explicitly modeling the trade-off between predictive performance and classifier diversity, which is crucial for achieving robustness and reproducibility in heterogeneous ECG datasets [19–21].

Recent studies have pointed out that ensemble diversity is crucial for improving the stability of classification, reducing overfitting, and increasing the detectability of minority irregular patterns [16,19–21].

Multi-objective optimization approaches, such as Multi-Objective Particle Swarm Optimization (MOPSO), provide efficient mechanisms for exploring trade-offs between conflicting objectives, such as maximizing predictive performance while preserving classifier diversity [22,23]. More broadly, evolutionary and metaheuristic algorithms have been successfully employed to optimize learning models for medical-data classification [24]. In the medical domain, MOPSO-based frameworks have also been applied to feature selection and diagnostic tasks involving heterogeneous biomedical datasets [25–27]. However, its utility in constructing diversity-aware ECG ensembles for the detection of AFib mostly remains unexplored [16,28].

In this study, we propose MOE-ECG, a novel Multi-Objective Ensemble Fusion framework that formulates ensemble selection as a bi-objective optimization problem. Instead of employing a predefined ensemble structure, the proposed approach automatically identifies subsets of complementary classifiers from a heterogeneous pool of 22 models by jointly optimizing classification performance and diversity using MOPSO. This selection mechanism is further combined with Dempster–Shafer evidence theory to produce robust probabilistic decisions. Unlike conventional ensemble approaches that focus primarily on aggregation, the key contribution of this work lies in the diversity-aware selection strategy and its systematic cross-dataset evaluation protocol designed to ensure stability and reproducibility.

The remainder of this paper is organized as follows: Section 2 provides the methodology of the proposed AFib detection framework in addition to the detailed description of the datasets, features, and classifiers used in this work. The results of the proposed model are presented and discussed in Section 3, followed by a conclusion in Section 4.

2. Materials and methods

2.1. Dataset

In this study, three publicly available ECG databases from PhysioNet [29] were used to evaluate the proposed MOE-ECG framework. The MIT-BIH Atrial Fibrillation Database (AFDB) served as the training set. AFDB contains 25 long-term two-channel ECG recordings of approximately 10 h each, sampled at 250 Hz, with beat-by-beat annotations indicating AFib and non-AFib rhythms [30]. Two recordings lacking annotation files were excluded from the analysis. The Saitama Heart Database Atrial Fibrillation (SHDB-AF) was used for model validation [31,32]. SHDB-AF comprises 128 Holter ECG recordings collected from 122 unique patients, each with a duration of exactly 24 h. The dataset contain two leads sampled at 125 Hz. Expert-labeled AFib/Atrial Flutter annotations are available for 98 of the 128 recordings. Where the SHDB-AF reference annotation identifies AFib/Atrial Flutter jointly, the database-provided positive rhythm annotation was retained as the positive class for the validation analysis. This dataset-specific convention differs from the AF-specific rhythm labeling available in AFDB and LTAFDB and is considered when interpreting cross-dataset results. Independent generalization performance was assessed using the Long-Term Atrial Fibrillation Database (LTAFDB), which includes 84 two-channel 24-hour Holter ECG recordings sampled at 128 Hz [33]. Each record contains detailed beat annotations enabling precise identification of AFib and normal sinus rhythm. All three datasets provide reference annotations indicating the timing and type of each heartbeat. A summary of the datasets used in this study is presented in Table 1.

2.2. Data preprocessing

The RR-intervals (RRIs) were extracted from each ECG recording as the basis for subsequent feature computation. For the retrospective experiments reported here, the expert beat annotations supplied with the databases were used as the primary R-peak locations, and the PhysioNet WFDB Toolbox [34] was used to access and process the corresponding annotation files. Accordingly, the reported classification results assess the framework under reliable beat localization; fully automated R-peak detection from noisy wearable ECG is outside the scope of the present evaluation. To maintain consistency across all datasets, only a single ECG lead was utilized from each record. The Modified Limb Lead II (MLII) was adopted for AFDB since it is the standard monitoring lead in PhysioNet arrhythmia studies and provides high R-wave amplitude for reliable R-peak detection. The CC5 lead was used for SHDB-AF due to its strong ventricular morphology and favorable signal-to-noise characteristics in Holter monitoring. For LTAFDB, the first Holter channel was selected as the primary acquisition channel to ensure reproducibility across subjects. All ECG recordings were resampled to a common sampling frequency of 250 Hz to ensure uniformity of R-peak timing across datasets. Let r_i denote the sample location of the ith R-peak and f_s the sampling frequency. The corresponding RR interval is defined as

$$RRI_i = \frac{r_{i+1} - r_i}{f_s}. \tag{1}$$

Each recording was segmented into windows containing $N \in \{20, 60, 120\}$ consecutive RRIs. Consecutive windows were generated with 50% overlap, corresponding to a stride of $N/2$. This value was selected as a compromise between temporal continuity and sample redundancy: each successive window retains half of the preceding rhythm context while introducing 50% new RR intervals. This improves coverage of rhythm transitions compared with non-overlapping windows without introducing the substantially higher redundancy associated with smaller strides. These window lengths approximately correspond to short (~15–20 s), medium (~45–60 s), and longer (~90–120 s) cardiac contexts at typical heart rates, which enables a trade-off between rapid AF detection and statistically stable variability estimation. Multi-scale

Table 1
Summary of the three ECG datasets used in this study.

Dataset	Record length (hrs)	Subjects #	Sampling rate (Hz)	AFib rhythms	AFib length (hrs)
AFDB	~10	23	250	291	~39
SHDB-AF	~24	122	125	809	~1459
LTAFDB	~24	84	128	7339	~1031

Table 2
Number of rhythms per class for different segment lengths among the three datasets.

Dataset	$RRI = 20$		$RRI = 60$		$RRI = 120$	
	Normal	AFib	Normal	AFib	Normal	AFib
AFDB	45432 (64.1%)	25424 (35.9%)	22739 (64.2%)	12689 (35.8%)	7578 (64.2%)	4231 (35.8%)
SHDB-AF	382004 (75.7%)	122737 (24.3%)	127319 (75.7%)	40928 (24.3%)	63653 (75.7%)	20470 (24.3%)
LTAFDB	291375 (46.7%)	332720 (53.3%)	145844 (46.7%)	166203 (53.3%)	48667 (46.8%)	55348 (53.2%)

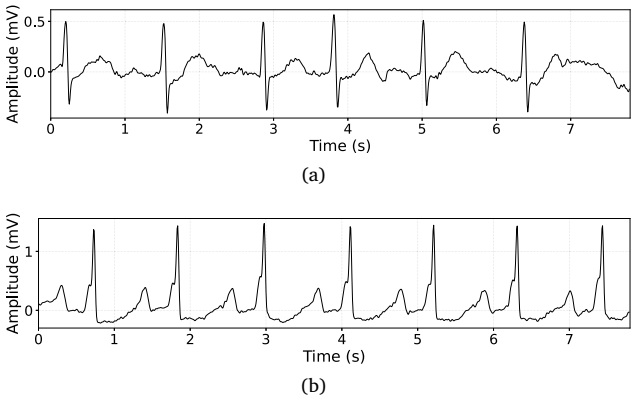


Fig. 1. An example of (a) AFib and (b) NSR segments. These are excerpts from the AFDB dataset. The y-axis represents the ECG amplitude in mV and the x-axis shows the time in seconds.

segmentation is commonly adopted in RR/HRV-based arrhythmia analysis to balance responsiveness and robustness [35–37]. For a window of N intervals, the window label was determined using the majority rule

$$y_w = \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^{N} a_i \geq 0.5\right), \tag{2}$$

where $a_i = 1$ denotes an AFib-labeled interval and $a_i = 0$ otherwise. Thus, a segment was labeled as AFib when at least 50% of its constituent RRIs were annotated as AFib; otherwise, it was labeled as Normal. For each segment length, the number of segments belonging to Normal and AFib classes was computed for the three datasets. These statistics are summarized in Table 2. As expected, AFDB exhibited moderate class imbalance, while SHDB-AF and LTAFDB yielded a more balanced distribution of AFib and Normal segments due to their longer continuous Holter recordings. Example RRI patterns for Normal sinus rhythm (NSR) and AFib rhythms are shown in Fig. 1, where the first exhibits regular and stable RR-intervals, whereas the latter displays beat-to-beat irregularity.

2.3. Feature extraction and selection

A total of fifteen RRI-based features were extracted for each segment to represent the statistical, temporal, and nonlinear properties of heart rate dynamics. The standard statistical features include mean RRI (Mean_RR), minimum RRI (Min), and maximum RRI (Max), which describe the general heart rate level and its extreme fluctuations [35, 38]. To assess variability, several time-domain measures were used. These included the standard deviation of RR intervals (SD_RR), the root mean square of successive differences (RMSSD), and the normalized root mean square of successive differences (nRMSSD), which capture both short-term and long-term beat-to-beat variability, often elevated in AFib [39,40]. Other HRV features, such as NN50, the number of successive RRI differences greater than 50 ms, and pNN50, the percentage of such differences relative to the total number of intervals were included to count and normalize the number of successive RRI differences greater than 50 ms, thus providing indications of abrupt irregularity in ventricular response [38]. For distributional spread, both mean absolute deviation (MeanAD) and the more robust median absolute deviation (MedianAD) were included [41]. Complexity and randomness in the RRI series are quantified using Shannon entropy [42]. Two ratio-based measures, the coefficient of variation of RR intervals (CVI) and the coefficient of variation of successive differences (CVS), both normalize variability with respect to mean levels [43]. The nonlinear dynamics are incorporated by the SD2/SD1 ratio derived from the Poincaré plot, which reflects the balance between long-term and short-term variability [44]. The average variation rate (AVRR) measures the mean magnitude of absolute successive changes in RRIs, which captures the rapid transitions characteristic of AFib [45].

These fifteen features together provide a multidimensional representation of heart rhythm characteristics, including variability, irregularity, and complexity, which are features known to be different in NSR and AFib. All RRI-based features extracted, with their short definitions and corresponding mathematical formulations, are summarized in detail in Table 3.

The present study intentionally focuses on RRI-derived features because beat-to-beat irregularity provides a compact representation of the ventricular rhythm characteristics associated with AFib. RRI-based representations are also less dependent on ECG amplitude scaling and lead-specific morphology, which is advantageous in a cross-dataset setting with heterogeneous acquisition conditions. However, this representation does not capture complementary morphological information such as atrial activity, P-wave characteristics, or QRS morphology.

Fig. 2 illustrates the Pearson pairwise correlation coefficients among all extracted features. The corresponding scatter plots for each feature pair are included, where the slope of the least-squares fitted line is equal to the correlation coefficient. Marginal histograms along the diagonal show the distribution of each normalized feature. The asterisks denote the statistical significance of the correlations: the larger the absolute value of the coefficient and the greater the number of asterisks, the stronger and more significant the relationship between the variables. The figure also displays each feature's individual association with the *Class* label (NSR vs. AFib) as well as the degree of inter-feature redundancy.

Several strongly correlated clusters appear, most prominently a short-term variability group consisting of RMSSD, nRMSSD, NN50, pNN50, MeanAD, and MedianAD, all exhibiting very high mutual correlations ($|r| > 0.85$). Likewise, SD_RR and Min form a highly collinear pair, and Entropy and CVI show substantial shared information.

Table 3
Summary of the extracted RRI-based features used for AFib detection.

Feature	Short definition	Mathematical formula
Mean_RR	Average time between successive RR intervals.	$Mean_RR = \frac{1}{N}\sum_{i=1}^{N} RRI_i$
SD_RR	Overall variability of RR intervals around the mean.	$SD_RR = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RRI_i - Mean_RR)^2}$
RMSSD	Beat-to-beat variability using successive differences.	$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RRI_{i+1} - RRI_i)^2}$
nRMSSD	RMSSD normalized by the mean RR interval.	$nRMSSD = \frac{RMSSD}{Mean_RR}$
Min	Shortest RR interval within the segment.	$Min = \min(RRI_1, RRI_2, \ldots, RRI_N)$
Max	Longest RR interval within the segment.	$Max = \max(RRI_1, RRI_2, \ldots, RRI_N)$
NN50	Count of successive RR differences greater than 50 ms.	$NN50 = \sum_{i=1}^{N-1} \mathbf{1}\{\lvert RRI_{i+1} - RRI_i\rvert > 50\,\text{ms}\}$
pNN50	Proportion of successive RR differences greater than 50 ms.	$pNN50 = \frac{NN50}{N-1}$
Entropy	Shannon entropy of the RR interval distribution.	$Entropy = -\sum_{j=1}^{K} p_j \log(p_j)$
MeanAD	Average absolute deviation of RRIs from their mean.	$MeanAD = \frac{1}{N}\sum_{i=1}^{N} \lvert RRI_i - Mean_RR\rvert$
MedianAD	Robust variability using deviation from median RRI.	$MedianAD = median(\lvert RRI_i - \tilde{RRI}\rvert)$
CVI	Relative variability: SD normalized by mean RRI.	$CVI = \frac{SD_RR}{Mean_RR}$
CVS	Coefficient of variation of successive RRI differences.	$CVS = \frac{SD(\Delta RRI)}{Mean(\Delta RRI)}$
SD2/SD1 ratio	Poincaré-based long-/short-term variability ratio.	$SD1 = \sqrt{\frac{1}{2}}SD(\Delta RRI),\ SD2 = \sqrt{2SD_RR^2 - SD1^2}$
AVRR	Average absolute change between successive RR intervals.	$AVRR = \frac{1}{N-1}\sum_{i=1}^{N-1} \lvert RRI_{i+1} - RRI_i\rvert$

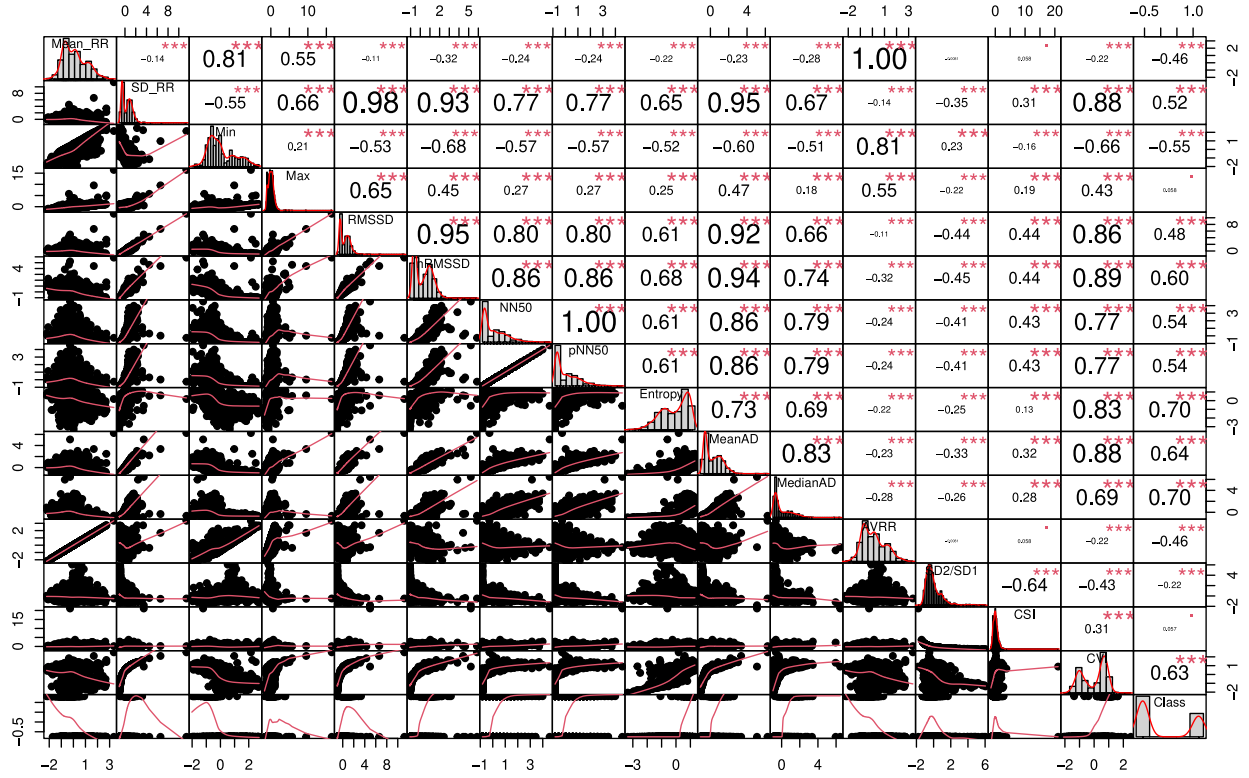


Fig. 2. The Pearson pairwise correlation coefficients are displayed together with the scatter plots of each feature pair and the marginal histograms along the diagonal. The slope of the least-squares fitted line in each scatter plot equals the corresponding correlation coefficient. Larger absolute values and a greater number of asterisks indicate more statistically significant correlations between feature pairs. This visualization highlights strongly correlated feature clusters, which were subsequently pruned to reduce redundancy during feature selection.

To reduce redundant physiological information, a two-stage feature selection strategy was applied. First, features were ranked according to the absolute Pearson correlation between each feature and the *Class* label. Second, redundancy pruning was performed by excluding any feature that had $|r| > 0.85$ with an already selected, higher-ranked feature. More formally, the relevance score of feature x_j was defined as $R_j = |\rho(x_j, y)|$. Starting from the highest-ranked feature, x_j was retained only if $|\rho(x_j, x_k)| \leq 0.85$ for every previously retained higher-ranked feature x_k. All relevance and redundancy calculations used to determine the feature subset were performed within the AFDB training-stage data. Once selected, the resulting feature subset was fixed and applied unchanged to SHDB-AF and LTAFDB; neither external dataset was used to redefine the selected features. This combined relevance–redundancy procedure resulted in a compact and informative subset of ten features, which are MedianAD, Entropy, CVI, MeanAD, nRMSSD, NN50, RMSSD, SD_RR, Mean_RR, and AVRR. Together, these features capture the key temporal, statistical, and nonlinear characteristics of AFib while minimizing multicollinearity and preserving model interpretability.

Fig. 3 shows the distribution of the selected ten RRI-based features for NSR and AFib segments by log-scaled boxplots. It can be seen

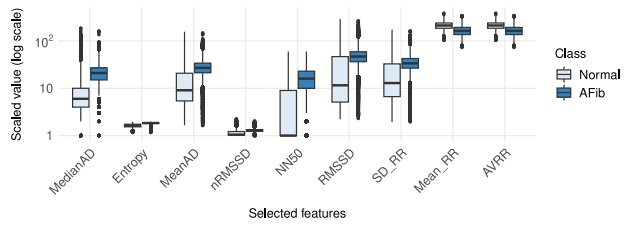
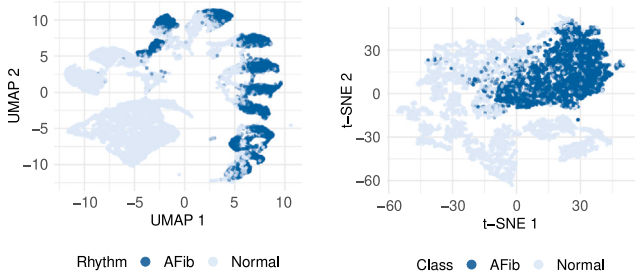


Fig. 3. Log-scaled boxplots of the selected ten RRI-based features for Normal and AFib segments. Clear separation between the two groups is visible for several features, particularly MedianAD, MeanAD, RMSSD, nRMSSD, NN50, and AVRR, which all show higher values in AFib due to increased beat-to-beat irregularity. Nonlinear measures such as Entropy and CVI also exhibit elevated values and wider dispersion in AFib, reflecting the greater complexity of arrhythmic RRI dynamics. Mean_RR shows a modest decrease for AFib, consistent with faster ventricular response. Overall, the distributions demonstrate that the selected features capture distinct and complementary AF-related characteristics suitable for classification.



(a) UMAP embedding of the selected features.

(b) t-SNE embedding of the selected features.

Fig. 4. Low-dimensional visualization of the ten selected RRI-based features. (a) UMAP and (b) t-SNE both show that AFib and Normal segments form distinguishable patterns, with AFib samples exhibiting more dispersed and heterogeneous structure due to their irregular RRI dynamics. These visualizations demonstrate that the selected features capture meaningful discriminative information suitable for AFib detection.

that several features, including MedianAD, MeanAD, RMSSD, nRMSSD, NN50, and AVRR, are higher in AFib due to the increased beat-to-beat irregularity in arrhythmic rhythm. Nonlinear measures of Entropy and CVI have values with higher spread in AFib due to the increased complexity of RRI dynamics. Mean_RR slightly shifts downward for AFib, consistent with faster ventricular response. Overall, the boxplots support the appropriateness of the selected features for classification, indicating that they capture complementary and distinctive signatures of AFib.

To further assess the discriminative power of the selected ten RRI-based features, we displayed their low-dimensional embeddings using Uniform Manifold Approximation and Projection (UMAP) [46] and t-distributed Stochastic Neighbor Embedding (t-SNE) [47]. As illustrated in 4(a) and 4(b), both methods suggest that AFib and NSR segments show a clear tendency to form separable structures, while AFib samples are more dispersed and heterogeneous due to their higher beat-to-beat irregularity. Although there is some overlap, which is expected when projecting such complex physiological dynamics into two dimensions, these embeddings confirm that the selected features capture meaningful structure and provide strong discriminative cues for distinguishing both classes.

2.4. Multi-objective ensemble classifier selection

To design a robust and accurate AFib detection system, we employed a multi-objective ensemble learning strategy based on a heterogeneous set of 22 classifiers and optimized them simultaneously for predictive performance and classifier diversity. AFib produces nonlinear and irregular fluctuations in the RRI series; therefore, using a diverse model pool helps capture complementary patterns that a single classifier family cannot learn effectively. This motivation is well-supported by ensemble learning theory [48–50].

Twenty-two classifiers were included, spanning linear, probabilistic, discriminant, distance-based, tree-based, boosting, bagging, modern gradient boosting (XGBoost, LightGBM), TabNet [51], and four neural networks (MLP, DeepNet, VGG, ResNet). Each classifier was tuned using a grid search technique [52,53] with stratified 5-fold cross-validation. This internal cross-validation was used only for hyperparameter optimization within AFDB and was not interpreted as the estimate of external generalization performance. Because overlapping windows introduce dependence between neighboring segments from the same recording, this internal tuning result is kept distinct from the subsequent cross-dataset evaluation. Ensemble selection and final testing were performed on the separate SHDB-AF and LTAFDB databases, respectively. The full set of hyperparameter ranges is summarized in Table 4. The best-performing configuration based on validation F_1-score was preserved for the ensemble selection stage.

After hyperparameter optimization, classifier selection was formulated as a bi-objective optimization problem. The selected ensemble should both maximize classification performance and maximize classifier diversity.

Let $\mathbf{s} = [s_1, s_2, \ldots, s_L]$ be the binary classifier selection vector, where $s_i = 1$ indicates that the ith classifier is included in the ensemble and $s_i = 0$ otherwise, and F_1 is the computed F_1-score of the ensemble composed of the selected classifiers. The ensemble performance objective is defined as the maximization of the validation F_1-score:

$$\max_{\mathbf{s}} \; f_1(\mathbf{s}) = F_1(\mathbf{s}). \tag{3}$$

There are two different approaches to measure the diversity of the selected classifiers; pairwise and non-pairwise [48]. In this paper, a pairwise disagreement among classifiers is computed [54]. For selected classifiers in $\mathbf{s}$, let $P^{(i)} \in \{0,1\}^N$ denote the hard prediction vector of classifier i, and let $L = \sum_i s_i$ be the total number of classifiers selected in the ensemble.

$$d_{ij} = 1 - \frac{1}{n}\sum_{r=1}^{n} \mathbf{1}\left(P_r^{(i)} = P_r^{(j)}\right), \tag{4}$$

Then, the overall ensemble diversity (average pairwise disagreement) is:

$$f_2(\mathbf{s}) = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} d_{ij}, \qquad L \geq 2. \tag{5}$$

2.5. Pareto optimality in multi-objective optimization

Pareto optimality is a fundamental concept in multi-objective optimization and forms the foundation for identifying the optimal trade-offs between conflicting design goals [55,56]. Within the scope of the proposed ensemble learning framework, performance and diversity are two such objectives that cannot be fully optimized at the same time without making some level of compromise. It is, therefore, essential to introduce the concept of Pareto optimality before detailing the MOPSO-based ensemble selection process.

Mathematically, a general multi-objective optimization problem with P objective functions and N inequality constraints can be formulated as follows [55,57]:

$$\min_{\mathbf{x}} \; \boldsymbol{H}(\mathbf{x}) = \left[h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_P(\mathbf{x})\right]^T, \tag{6}$$

subject to

$$g_n(\mathbf{x}) \leq 0, \qquad n = 1, 2, \ldots, N, \tag{7}$$

where $\boldsymbol{x}$ is an m-dimensional decision vector taken from a feasible region. The functions $\boldsymbol{H}(\boldsymbol{x})$ represent the P conflicting objectives, while

Table 4
Hyperparameter ranges explored for each classifier during grid search.

Classifier	Hyperparameters
MLP	Number of hidden layers = [1, 2, 3]; Number of units = [32, 64, 128]; Activation function = [relu, elu]; Optimizer = [adam,rmsprop]; Batch size = [32, 64]; Number of epochs = [10, 20].
Deep Neural Network	Number of layers = [2, 3, 4]; Number of units = [128, 256]; Activation function = [relu, elu]; Optimizer = [adam,rmsprop]; Batch size = [32, 64]; Number of epochs = [10, 20].
VGG	Convolution block sizes = [[32, 64],[64, 128]]; Dense units = [256, 512]; Dropout rate = [0.3, 0.4]; Activation function = [relu]; Optimizer = [adam]; Batch size = [32]; Number of epochs = [10].
ResNet	Number of filters = [32, 64]; Number of residual blocks = [2, 3]; Dense units = [64, 128]; Dropout rate = [0.2, 0.3]; Activation function = [relu]; Optimizer = [adam]; Batch size = [32]; Number of epochs = [10].
TabNet	Feature dimension (n_d) = [8, 16, 32]; Attention dimension (n_a) = [8, 16, 32]; Number of decision steps = [3, 5]; Gamma (relaxation) = [1.0, 1.5]; Sparsity regularization = [1e−3, 1e−4]; Momentum = [0.02]; Mask type = [sparsemax].
Logistic Regression	Regularization strength C = [0.001, 0.01, 0.1, 1]; Penalty type = [l1, l2]; Solver = [liblinear, saga]; Maximum number of iterations = [200, 500].
Random Forest	Number of estimators = [100, 200, 300]; Maximum depth = [None,10, 20, 30]; Minimum samples split = [2, 5, 10]; Minimum samples per leaf = [1, 2, 4]; Maximum number of features = [sqrt,log2].
XGBoost	Number of estimators = [100, 200]; Maximum depth = [3, 6, 9]; Learning rate = [0.01, 0.1, 0.3]; Subsample ratio = [0.8, 1.0]; Column subsample (per tree) = [0.8, 1.0]; Gamma (tree regularization) = [0, 0.1, 0.2].
LightGBM	Number of estimators = [100, 200]; Number of leaves = [31, 63, 127]; Learning rate = [0.01, 0.1]; Subsample ratio = [0.8, 1.0]; Feature subsample ratio = [0.8, 1.0]; Minimum child samples = [20, 50].
Decision Tree	Maximum depth = [None,10, 20, 30]; Minimum samples split = [2, 5, 10]; Minimum samples per leaf = [1, 2, 4]; Splitting criterion = [gini,entropy].
K-Nearest Neighbors	Number of neighbors = [3, 5, 7, 9]; Weighting scheme = [uniform, distance]; Distance metric = [euclidean, manhattan]; Minkowski power p = [1, 2].
GaussianNB	Variance smoothing = [1e−9].
BernoulliNB	Smoothing parameter α = [0.1, 0.5, 1.0]; Binarization threshold = [0.0, 0.5].
LinearSVC	Regularization strength C = [0.1, 1, 10]; Loss function = [hinge, squared_hinge]; Maximum number of iterations = [1000, 2000].
RidgeClassifier	Regularization strength α = [0.1, 1.0, 10.0]; Solver = [auto, cholesky]; Maximum number of iterations = [None, 500].
SGDClassifier	Loss function = [hinge, log_loss,modified_huber]; Penalty = [l2, l1, elasticnet]; Regularization parameter α = [1e−4, 1e−3, 1e-2]; Maximum number of iterations = [1000]; Tolerance = [1e−3].
ExtraTrees	Number of estimators = [100, 200]; Maximum depth = [None, 10, 20]; Minimum samples split = [2, 5]; Minimum samples per leaf = [1, 2]; Maximum number of features = [sqrt, log2].
AdaBoost	Number of estimators = [50, 100]; Learning rate = [0.1, 1.0]; Boosting algorithm = [SAMME, SAMME.R].
Gradient Boosting	Number of estimators = [100, 200]; Learning rate = [0.01, 0.1]; Maximum depth = [3, 5]; Minimum samples split = [2, 5]; Minimum samples per leaf = [1, 2].
Bagging	Number of estimators = [10, 20, 50]; Maximum samples ratio = [0.5, 0.7, 1.0]; Maximum features ratio = [0.5, 0.7, 1.0].
LDA	Solver = [svd, lsqr, eigen]; Shrinkage = [None, auto]; Tolerance = [1e−4, 1e−3].
QDA	Regularization parameter = [0, 0.1, 0.2]; Tolerance = [1e−4, 1e−3].

$g_n(\boldsymbol{x})$ are constraint functions. In multi-objective optimization, the objective functions may be linear or nonlinear, convex or non-convex.

In almost all real-world multi-objective optimization problems, including classifier ensemble selection, the objectives are mutually conflicting. As a result, one cannot optimize all objectives simultaneously. Instead, the goal is to identify solutions that offer the best possible trade-off. These solutions are referred to as *Pareto optimal* or *non-dominated* solutions.

To formally define these concepts, consider two objective vectors $\boldsymbol{u} = (u_1, u_2, \ldots, u_P)$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_P)$ associated with two feasible candidate solutions. For a minimization problem, the concept of dominance is defined as follows [55,57]:

- *Dominance:* A vector $\boldsymbol{u}$ is said to dominate $\boldsymbol{v}$ (denoted $\boldsymbol{u} \prec \boldsymbol{v}$) if and only if $\forall\, i \in \{1, 2, \ldots, L\},\ u_i \leq v_i \wedge\ \exists\, i \in \{1, 2, \ldots, L\} : u_i < v_i$.
- *Pareto optimality:* A candidate solution $\boldsymbol{x}^\star \in \Theta$ is considered Pareto optimal if, within the feasible set Θ, there exists no alternative solution x such that the objective vector $\boldsymbol{H}(\boldsymbol{x}) = [h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_P(\boldsymbol{x})]$ is strictly superior to $\boldsymbol{H}(\boldsymbol{x}^\star) = [h_1(\boldsymbol{x}^\star), h_2(\boldsymbol{x}^\star), \ldots, h_P(\boldsymbol{x}^\star)]$. in all objectives and strictly better in at least one of them. In other words, $\boldsymbol{x}^\star$ is non-dominated with respect to all other feasible solutions.

The set of all non-dominated solutions forms the *Pareto front*. In the proposed ensemble learning framework, each point on the Pareto front represents a specific classifier subset characterized by a trade-off between ensemble performance and classifier diversity. Using MOPSO [56], the algorithm maintains an external repository of such non-dominated solutions, and the final ensemble is selected from this set based on validation F_1-score. This ensures that the chosen ensemble is both accurate and diverse while adhering to the principles of multi-objective optimality.

2.6. Multi-objective PSO

To determine the best subset of the classifiers for the proposed ensemble, we use a multi-objective particle swarm optimization algorithm. MOPSO extends the classical PSO method, which is used to optimize a single objective. In contrast, MOPSO is designed to handle multiple, possibly conflicting objectives. In our case, the objectives correspond to the following: (i) maximizing predictive performance of the ensemble and (ii) enhancing classifier diversity. Optimizing these two quantities jointly will ensure that the final ensemble is both accurate and composed of diverse models.

In the standard PSO algorithm, each particle represents a candidate solution, whose position is iteratively updated. The velocity update at iteration k is governed by three components: 1) an inertia term encouraging continuation of previous motion, 3) a cognitive (nostalgia) term pulling particles toward their own best historical solution, and 3) a social (global) term attracting them toward the swarm leader or *global*

best. These update rules are formulated as [58]:

$$\begin{aligned} \boldsymbol{v}_i(k) = & \overbrace{\omega \cdot \boldsymbol{v}_i(k-1)}^{\text{inertia term}} \\ & + \overbrace{c_1 \cdot r_1 \left(\boldsymbol{x}_{pbest,i} - \boldsymbol{x}_i(k)\right)}^{\text{nostalgia term}} \\ & + \overbrace{c_2 \cdot r_2 \left(\boldsymbol{x}_{gbest,i} - \boldsymbol{x}_i(k)\right)}^{\text{global term}}. \end{aligned} \tag{8}$$

$$\boldsymbol{x}_i(k) = \boldsymbol{x}_i(k-1) + \boldsymbol{v}_i(k), \tag{9}$$

where ω is the inertia weight, $\mathbf{x}_{pbest,i}$ denotes the personal best position of particle i, $\mathbf{x}_{gbest,i}$ denotes the position of the swarm leader, and r_1 and r_2 are uniformly distributed random variables in $[0, 1]$, and c_1 and c_2 represent the learning coefficients that determine the relative contribution of the cognitive (*personal best*) and social (*global best*) components. In this work, we employ parameter values consistent with established recommendations in earlier studies, which are $c_1 = c_2 = 2$ and $\omega = 0.8$ [59].

Unlike classical PSO, the MOPSO algorithm does not rely on a single global best strategy. Instead, it maintains an external repository (or *archive*) of non-dominated solutions, forming an approximation of the Pareto-optimal set [60]. When selecting a leader for a particle, MOPSO samples from this repository rather than using a single best position since they are all non-dominated solutions. To improve the spread of solutions across the Pareto front, a region-based leader selection mechanism is adopted in which the objective space is divided into subregions [22,56,61]. Leaders from sparsely populated regions are given higher selection probability, which encourages search diversity and promotes the discovery of non-dominated solutions as *global best* candidates.

We also apply a mutation operator, as recommended in [22], to prevent premature convergence and to explore new areas of the search space. After each update, the dominated solutions are removed from the repository to ensure that only Pareto-efficient candidates remain. To prevent the repository from exceeding its predefined capacity, only the set of non-dominated leaders is retained, and the occupancy of the subregions is updated accordingly. Before executing the MOPSO algorithm, several key parameters must be specified. These include:

- *nPop*; population size
- *MaxIt*; maximum number of iterations
- *nRep*; repository (archive) size
- μ; mutation rate
- β; leader-selection pressure
- *nRegion*; number of subregions in the objective space
- c_1 and c_2; learning coefficients
- ω; inertia weight

The overall MOPSO procedure used in this study are provide in Algorithm 1.

Algorithm 1: MOPSO algorithm

Input : MOPSO parameters $(nPop, MaxIt, nRep, \mu, \beta, nRegion, c_1, c_2, \omega)$
Output: Pareto optimal set (non-dominated solutions)

```
Set the values of MOPSO parameters
Initialize the position and the velocity of the swarm
Evaluate objective values on initialized particles
Select non-dominated solutions as leader g_best
for it ← 0 to MaxIt do
    Update the archive-based leader selection for the new iteration
    for n ← 0 to nPop do
        Select a leader for particle n
        Update the velocity and position of particle n
        Apply mutation on particle n's position
    Evaluate objective values
    Add non-dominated particles to the repository
    Determine domination of new repository members
    Keep only non-dominated members in the repository
    Remove members from occupied subregions if repository is full (nRep)
Report Pareto optimal set (non-dominated solutions)
```

2.7. Dempster–Shafer combination rule

To fuse the outputs of the classifier ensemble selected by the MOPSO algorithm, we adopt the Dempster–Shafer Theory (DST) [62], which provides a rigorous framework for combining independent sources of evidence and quantifying uncertainty. Unlike traditional probabilistic fusion, DST allows each classifier to contribute a degree of belief to individual hypotheses as well as to sets of hypotheses, making it especially suitable for medical decision systems where uncertainty and disagreement between classifiers may occur.

DST is defined over a universal hypothesis set $X = \{\text{Normal}, \text{AFib}\}$. For any subset $A \subseteq X$, the mass function $m : 2^X \to [0, 1]$ assigns a basic probability value according to:

$$m(\emptyset) = 0, \tag{10}$$

$$\sum_{A \subseteq X} m(A) = 1,$$

where $m(A)$ represents the support committed exactly to subset A. In this study, the normalized two-class confidence scores produced by each selected classifier are used directly as the basic mass assignments for the two singleton hypotheses. For classifier i, let $c_{i,N}$ and $c_{i,A}$ denote the normalized confidence scores for Normal and AFib, respectively, with $c_{i,N} + c_{i,A} = 1$. The corresponding masses are

$$\begin{aligned} & m_i(\{\text{Normal}\}) = c_{i,N}, \\ & m_i(\{\text{AFib}\}) = c_{i,A}, \qquad m_i(X) = 0. \end{aligned} \tag{11}$$

Thus, although DST can assign explicit mass to the composite uncertainty set, the present binary implementation places the available mass on the two singleton hypotheses. The normalized outputs are used as fusion confidence scores and should not be interpreted as independently calibrated clinical probabilities.

DST combines independent mass functions m_1 and m_2 using the Dempster–Shafer combination rule:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \emptyset} m_1(B)\, m_2(C), \tag{12}$$

$$\text{where} \quad K = \sum_{B \cap C = \emptyset} m_1(B)\, m_2(C),$$

is the conflict coefficient quantifying disagreement between the two sources. The larger the value of K the more inconsistent the evidence. The implementation follows the standard Dempster normalization for $K < 1$ and does not introduce an additional conflict-resolution heuristic. For illustration, if two classifiers assign AFib confidences of 0.80 and 0.70, then $K = (0.80 \times 0.30) + (0.20 \times 0.70) = 0.38$. The combined AFib and Normal masses are therefore $0.56/(1 - 0.38) \approx 0.903$ and $0.06/(1 - 0.38) \approx 0.097$, respectively.

The combination rule is applied iteratively to fuse the mass functions of all selected classifiers, producing a final aggregated mass distribution. From this, both the belief (Bel) and plausibility (Pl) measures can be derived, yielding lower and upper bounds on the confidence associated with each class. The final combined mass yields a single, robust estimate of the likelihood of AFib versus normal rhythm.

In the present implementation, DST provides a principled mechanism for combining the normalized evidence of the selected classifiers while accounting for inter-classifier conflict through the Dempster normalization rule.

2.8. MOE-ECG framework for AFib detection

In this subsection, we present the workflow of the proposed MOE-ECG method for robust AFib detection. The framework utilizes multi-objective classifier selection and evidence fusion using DST. It consists

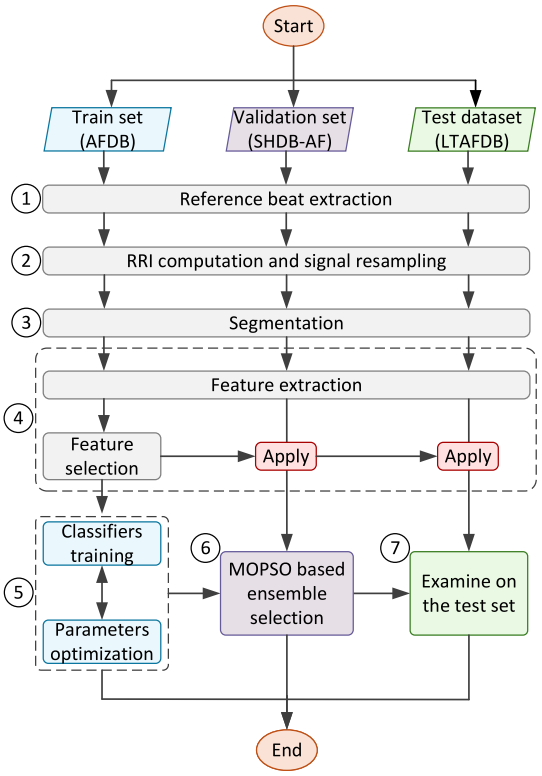


Fig. 5. Flowchart of the proposed MOE-ECG algorithm.

of several steps, which begins with preprocessing and segmentation of ECG signals, extraction of RRI–based features, selection of an informative and non-redundant feature subset, and finally, construction of a diverse and accurate ensemble using MOPSO. The flowchart of the proposed MOE-ECG approach is illustrated in Fig. 5, which is described in the following main steps:

1. *Reference beat extraction*: R-peak locations are obtained from the expert beat annotations provided with AFDB, SHDB-AF, and LTAFDB and are used as the basis for deriving the RRI sequences.
2. *RRI computation and signal resampling*: Each ECG recording is resampled to 250 Hz for consistency across the datasets. RRIs are computed as the time differences between consecutive R-peaks, which is used as the basis for rhythm characterization.
3. *Segmentation into 20, 60, and 120 RRIs*: The RRI sequences are segmented into windows of size 20, 60, and 120 with 50% overlap. Each segment is labeled as AFib or Normal using a majority rule threshold, where $\geq 50\%$ AFib RRIs define an AFib window.
4. *Feature extraction and selection*: The fifteen features, as described in Section 2.3 and given in Table 3, are extracted from the RRIs. The two-stage relevance–redundancy procedure is determined using AFDB training data, after which the selected ten-feature subset is applied unchanged to SHDB-AF and LTAFDB.
5. *Classifiers training and hyperparameter optimization*: A diverse pool of classifiers are trained and optimized using Grid Search and stratified 5-fold cross-validation technique on the training set (AFDB).
6. *Ensemble selection*: The MOPSO procedure is applied to identify accurate and diverse classifier subsets. All non-dominated solutions are evaluated on SHDB-AF, and the solution with the highest validation F_1-score is selected as the final ensemble.

Table 5
Confusion matrix.

	Predicted negative	Predicted positive
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

7. *Independent test evaluation*: The ensemble selected using SHDB-AF is evaluated on the independent LTAFDB test dataset. The outputs of the selected classifiers are combined using the Dempster–Shafer fusion rule described in Section 2.7.

3. Experimental validation

As described in Section 2, the proposed MOE-ECG framework was evaluated using three well-known ECG databases for AFib detection. The AFDB [30] was employed for training the individual classifiers and the SHDB-AF [31,32] were used to find the best ensemble of classifiers among the non-dominated solutions. To further assess the generalizability and robustness of the resulting ensemble, an additional independent dataset, the LTAFDB [33], was used for external testing.

Various standard metrics were derived from the confusion matrix, as given in Table 5, to quantitatively assess the performance of the proposed method. Each row in Table 5 corresponds to the model predictions and each column represents the ground-truth labels. Accuracy (Acc), sensitivity (Se), specificity (Sp), precision (Pr), false positive rate (FPR), and the F_1-score are the metrics that have been used in this study. These are formally defined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{13}$$

$$Se = \frac{TP}{TP + FN}, \tag{14}$$

$$Sp = \frac{TN}{TN + FP}, \tag{15}$$

$$Pr = \frac{TP}{TP + FP}, \tag{16}$$

$$FPR = \frac{FP}{FP + TN}, \tag{17}$$

$$F_1 = \frac{2\,Pr\,Se}{Pr + Se}, \tag{18}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The F_1-score is the harmonic mean of sensitivity and precision.

3.1. Classification performance of the proposed MOE-ECG algorithm

As illustrated in Fig. 6, the MOPSO algorithm produced a well-structured Pareto front, which consists of multiple non-dominated solutions shown in blue circles. Each of the non-dominated solutions represents a distinct balance between classifier performance (F_1-score) and diversity. Then, all non-dominated candidate ensembles were evaluated on the validation dataset (SHDB-AF) to examine their generalization capability beyond the training set. The non-dominated solution with the highest F_1-score on the SHDB-AF validation dataset, highlighted by the red star in Fig. 6, was selected as the final MOE-ECG ensemble; LTAFDB was not used in this selection. As reported in Fig. 6, this optimal solution includes these four base classifiers: TabNet, XGBoost, LightGBM, and Bagging. This created ensemble is used for presenting the performance evaluation on the unseen test data (LTAFDB) using 60-beat ECG segments.

To evaluate the performance of the proposed MOE-ECG ensemble in classifying short-term AFib, we tested its performance with input segments composed of 60 consecutive RRIs. This duration provides sufficient temporal context to capture irregular ventricular response

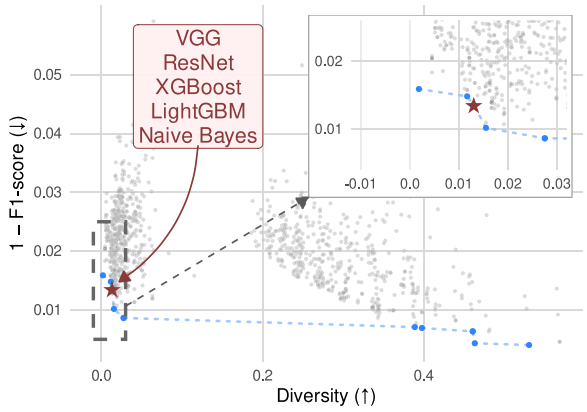


Fig. 6. Pareto front of the MOPSO based ensemble approach. Blue points denote non-dominated solutions, while the red star marks the selected ensemble achieving the best validation performance on SHDB-AF and used for testing on the test data (LTAFDB). The x-axis represents classifier diversity, where higher diversity indicates a more complementary ensemble, and the y-axis shows 1-F_1-score, where lower values correspond to better classification performance.

patterns while retaining a relatively short analysis window. Computational and device-level feasibility were not evaluated in the present study. The final ensemble from the MOPSO based classifier selection process was evaluated on the independent LTAFDB test set to measure its capability to distinguish between AFib and Normal rhythms.

To better assess the stability of the MOPSO-based selection mechanism, Fig. 7 illustrates the top-10 most frequently selected classifiers for each segment length (20, 60, and 120 RRIs). The figure highlights the dominant contributors (classifiers) to the Pareto optimal ensembles (blue circles in Fig. 6) across five independent optimization runs.

Boosting and tree-based methods such as XGBoost, Bagging, LightGBM, Decision Tree, and Random Forest, consistently rank among the most frequently selected classifiers across all segment lengths. This indicates that the proposed multi-objective ensemble model favors models that achieve strong F_1-score while maintaining structural diversity. On the other hand, these classifiers can efficiently capture the nonlinear and irregular characteristics of AFib cases.

Models such as QDA and k-NN appear intermittently, which suggests that complementary decision boundaries also contribute to the performance–diversity balance. In contrast, linear and naïve Bayes classifiers are rarely selected, which indicates their limited contribution under the proposed optimization criteria. Overall, Fig. 7 confirms that MOPSO consistently converges toward a compact and reproducible subset of high-performing and complementary classifiers, which shows the stability of the proposed ensemble strategy.

8(a) presents the ROC curve of the resulting ensemble, which exhibits excellent discrimination across a wide range of decision thresholds. The curve achieves an area under the curve (AUC) of well above 0.94, which indicates the probability outputs of the model provide strong separation between AFib and NSR. As given in the figure, several operating points, which correspond to thresholds of 0.7, 0.5, and 0.2, are highlighted on the ROC curve to illustrate how different cutoff values influence sensitivity and specificity. The pink shaded regions surrounding these points represent uncertainty bands derived from bootstrapped resampling. Narrow regions indicate stable performance, whereas wider bands suggest greater sensitivity to data fluctuations. Together, these visual cues provide additional context for interpreting the Youden J point (0.8071), which corresponds to the threshold that maximizes the combined sensitivity and specificity and serves as an optimal operating point for AFib detection.

In addition, to derive an operationally optimal threshold value, we computed the Kolmogorov–Smirnov (KS) statistic [63], that indicates the maximum separation of empirical cumulative distribution function (CDF) [64] of AFib and Normal prediction scores. In Fig. 8(b), the maximum KS value of 0.8046 is obtained at the threshold of 0.7803. This large KS distance confirms that the two score distributions are strongly separated and provides a principled threshold choice for binary classification.

The optimal decision threshold obtained from the ROC analysis (Fig. 8(a)) was determined using the Youden's J statistic, which is defined as $J = \mathrm{Se} + \mathrm{Sp} - 1$ [65]. As shown in Fig. 8(a), the Youden's J point occurred at a threshold equal to 0.8071, which represents the point that maximizes the balance between Se and Sp. This value is highly consistent with the threshold obtained from the KS analysis (Fig. 8(b)), which identifies the maximum class-separation at threshold of approximately 0.78. The close agreement between these two independently derived thresholds indicates that the ensemble classifier shows stable and robust separation between AFib and Normal rhythms, with minimal sensitivity to small shifts in the decision boundary.

To further explore the effect of thresholding, three versions of the confusion matrix plot were generated, each showing how the false prediction rate varies across possible score cutoffs/thresholds. Fig. 9 presents the confusion matrices for the three different thresholds of 0.5 (default), 0.8071 (Youden's J; Fig. 8(a)), and 0.7803 (KS analysis; Fig. 8(b)), respectively. These figures clearly demonstrate that thresholds near 0.78–0.81 provide the best trade-off between reducing false AFib detections and preserving AFib sensitivity.

Overall, the performance analysis on 60-beat ECG segments indicates strong class separability and stable behavior across the examined thresholds. Although the segment duration is relatively short, real-time suitability cannot be inferred because computational and device-level feasibility were not evaluated in the present study.

3.2. Comparison of different segment lengths

To investigate how the temporal context of the ECG affects atrial fibrillation detection, the MOE-ECG framework was evaluated using three segment lengths: 20, 60, and 120 heartbeats. Table 6 summarizes the performance of the proposed MOE-ECG model on the test set across the metrics defined in Eqs. (13) and (18). The metrics are reported as mean ± SD together with the 95% confidence intervals.

Overall, the results show a consistent performance improvement when increasing the segment length from 20 to 60 beats, followed by a performance plateau between 60 and 120 beats. From Table 6, the 60 beat segment length gives the most balanced performance across all metrics. For example, accuracy increases from 0.825 to 0.847 when moving from 20 to 60 beats and the F_1-score improves by more than 1.5%, which reflects more reliable classification between AFib and normal rhythms. Sensitivity shows a notable improvement as well, which increases from 0.951 at 20 beats to 0.982 at 60 beats.

On the other hand, increasing the window size to 120 beats does not lead to substantial improvement. For instance, accuracy, Se, and Sp remain nearly identical to those obtained with 60 beats, which confirms the hypothesis of performance saturation once sufficient temporal variability is captured. Precision shows only a marginal increase from 60 to 120 beats (i.e., from 0.697 to 0.705), which is consistent with the stable FPR across the longer segment lengths.

These findings indicate that the 20-beat segment length provides less temporal context, resulting in higher variance and reduced specificity. In contrast, 60-beat windows provide a useful balance between temporal context and segment duration, with little additional benefit observed at 120 beats for most metrics. The 60-beat configuration was therefore selected as the primary operating point for the subsequent analyses in this study.

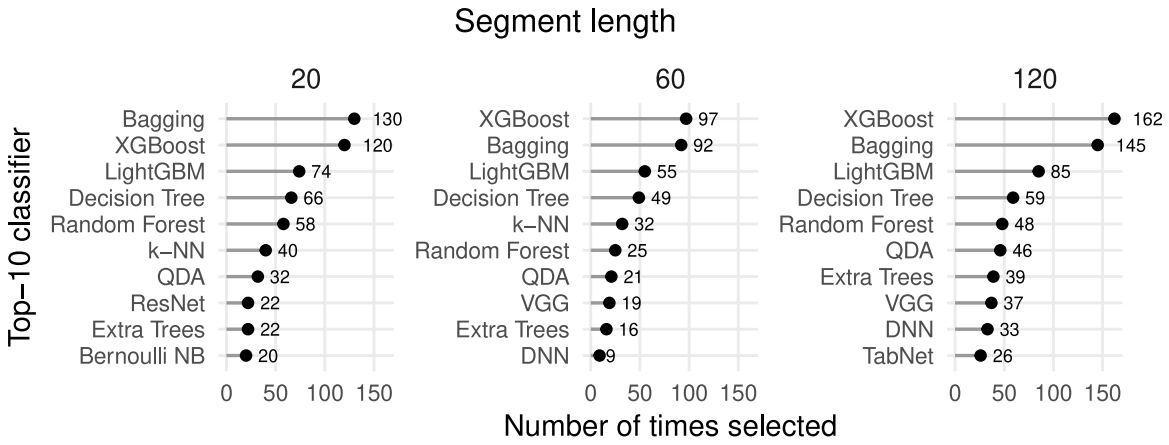
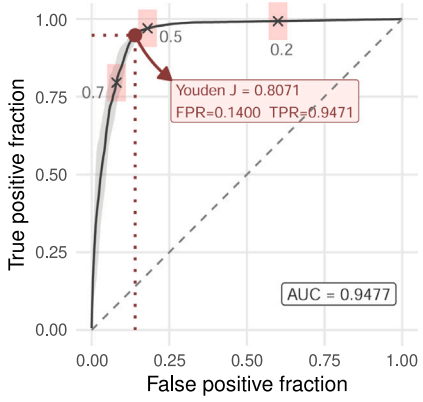
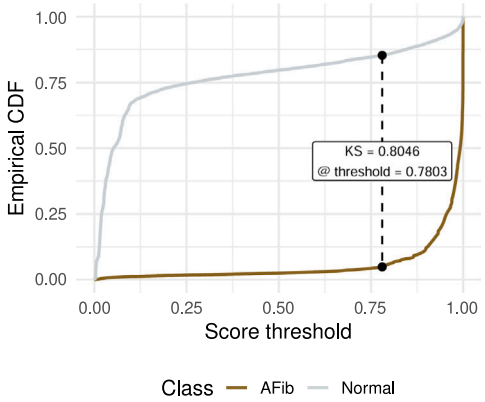


Fig. 7. Top-10 most frequently selected classifiers across five MOPSO optimization runs for segment lengths of 20, 60, and 120 RRIs. The lollipop plot highlights the dominant contributors to the Pareto-optimal solutions (blue circles in Fig. 6).

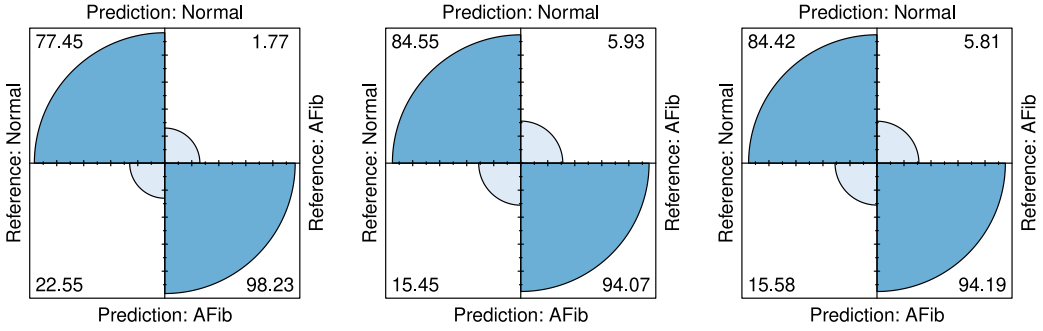


(a) ROC curve of the MOE-ECG ensemble for 60-beat ECG segments, demonstrating good discrimination between AFib and NSR (AUC = 0.9477).



(b) KS analysis showing maximum separation between the predicted score distributions of AFib and NSR. The peak KS value of 0.8046 occurs at a threshold of approximately 0.78.

Fig. 8. Performance evaluation of the MOE-ECG ensemble on 60-beat ECG segments. (a) ROC curve quantifying overall discrimination performance. (b) KS statistic identifying the optimal threshold that maximizes separation between AFib and NSR probability distributions. Both plots use harmonized axis scaling and grid style for visual consistency.



(a) Threshold = 0.50 (default). (b) Threshold = 0.8071 (Youden J). (c) Threshold = 0.7803 (KS).

Fig. 9. Confusion matrices of the MOE-ECG ensemble under three decision thresholds for AFib classification on 60-beat ECG segments. (a) Default threshold of 0.50. (b) Youden J optimal threshold of 0.8071, which maximizes the sum of sensitivity and specificity. (c) KS optimal threshold of 0.7803, which maximizes the separation between the predicted-score distributions of AFib and Normal rhythms. These comparisons highlight how threshold choice influences FP, FN, and overall classification performance.

Table 6
Comparison of MOE-ECG performance across three segment lengths on the test set (LTAFDB). Values are reported as mean (with SD and 95% CI in smaller font).

Metric	20 beats	60 beats	120 beats
Acc	0.825±0.036 [0.781, 0.870]	**0.847**±0.005 [0.841, 0.853]	0.843±0.008 [0.833, 0.853]
Se	0.951±0.012 [0.939, 0.967]	**0.982**±0.002 [0.979, 0.985]	0.979±0.003 [0.975, 0.983]
Sp	0.756±0.049 [0.695, 0.817]	**0.776**±0.008 [0.765, 0.786]	0.765±0.014 [0.748, 0.783]
Pr	0.684±0.044 [0.629, 0.738]	0.697±0.008 [0.688, 0.706]	**0.705**±0.013 [0.689, 0.721]
F_1-score	0.795±0.034 [0.753, 0.838]	0.815±0.004 [0.810, 0.821]	**0.820**±0.008 [0.810, 0.829]
FPR	0.244±0.049 [0.183, 0.305]	**0.224**±0.008 [0.214, 0.235]	0.235±0.014 [0.217, 0.252]

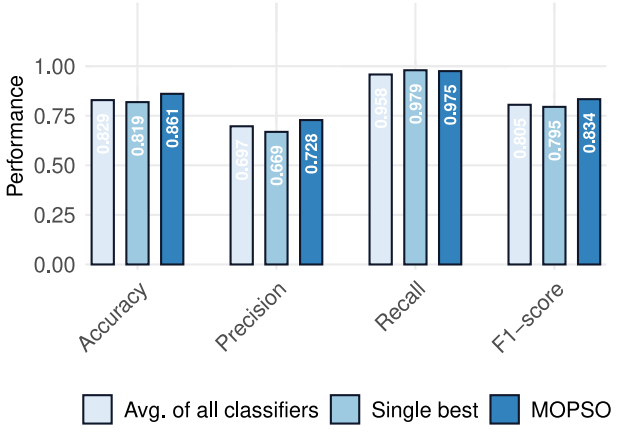


Fig. 10. Performance comparison between the proposed MOE-ECG ensemble, the average of all classifiers, and the single-best classifier on 60-beat ECG segments across five independent runs. Bars show the mean accuracy, precision, recall, and F_1-score. Given the limited number of independent runs, the differences are interpreted descriptively rather than through formal hypothesis testing.

3.3. Comparison with the average ensemble and the single best classifier

Fig. 10 presents a comparative analysis of the proposed MOE-ECG ensemble with two baseline approaches on the 60 beat ECG segments: (i) the average ensemble, which is obtained by computing the mean output across all individual classifiers, and (ii) the single best classifier, which is identified as the model with the highest F_1-score on the validation data and subsequently evaluated on the unseen LTAFDB test set. The figure reports the mean classification performance across five independent runs for four key metrics (i.e., *Acc*, *Pr*, *Se*, and F_1-score) for each of the three approaches.

Fig. 10 shows that MOE-ECG has higher mean accuracy, precision, and F_1-score than the two baseline approaches, whereas recall is similar and is marginally higher for the single-best classifier. This pattern suggests a favorable balance between sensitivity and precision for the selected ensemble, while the magnitude of the differences is interpreted descriptively because only five independent runs were available.

Given the limited number of independent runs ($n = 5$), the comparisons in Fig. 10 are interpreted descriptively rather than through formal hypothesis testing. The mean F_1-score was 0.834 for MOE-ECG, compared with 0.805 for the average ensemble and 0.795 for the single-best classifier, corresponding to absolute differences of 0.029 and 0.039, respectively. Accuracy and precision show the same directional pattern, while recall remains comparable across the three approaches. The values in Fig. 10 are intended as a within-run baseline comparison; the primary external-test performance summary is reported separately in Table 6.

Overall, the repeated-run comparison indicates that the diversity-aware selection and evidence-fusion strategy provides favorable mean accuracy, precision, and F_1-score relative to the two considered baselines, while maintaining comparable recall.

3.4. Performance comparison with state-of-the-art

Table 7 provides a contextual comparison of MOE-ECG with published AFib detection methods that report results on LTAFDB. The studies represent diverse modeling approaches, including LSTM-based recurrent networks, CNN and CNN–Transformer hybrids, residual and Transformer architectures, Sobolev-based learning models, and classical ensemble approaches such as AdaBoost. Because the training data, input durations, preprocessing, and evaluation protocols are not identical across studies, these results provide useful context but do not constitute a controlled head-to-head comparison.

The reported performances show considerable variation. The accuracy, sensitivity, and F_1-score values, if available, range from approximately 0.827 to 0.913, 0.657 to 0.959, and 0.740 to 0.915, respectively. As given in Table 7, models using Transformer blocks or Sobolev test-statistics tend to provide high sensitivity, while convolutional and recurrent architectures often achieve better accuracy but may be more affected by noise or domain shift when evaluated on unseen ECG recordings.

Because the studies summarized in Table 7 differ in training datasets, preprocessing procedures, input durations, model-development strategies, and evaluation protocols, the comparison should be interpreted as contextual rather than as a controlled head-to-head benchmark. Within this context, MOE-ECG achieved an accuracy, sensitivity, and F_1-score of 0.847, 0.982, and 0.815, respectively, on the independent LTAFDB test set. Its sensitivity of 0.982 is the highest among the methods included in the comparison, indicating a sensitivity-oriented operating profile with relatively few missed AFib episodes. Several competing approaches report higher accuracy and/or F_1-score, so the results should not be interpreted as uniformly superior performance across all metrics.

Overall, Table 7 indicates that MOE-ECG provides competitive external-test performance while emphasizing high sensitivity. Accordingly, this cross-study comparison is interpreted as contextual evidence of competitive performance rather than definitive evidence of superior robustness or generalization.

3.5. Limitations and future work

Several limitations should be considered when interpreting the present findings. First, MOE-ECG relies on RRI-derived characteristics and therefore does not exploit complementary ECG morphological information such as atrial activity, P-wave characteristics, or QRS morphology. Future work could investigate the integration of RRI-based features with morphology-based or learned ECG representations to determine whether these complementary sources of information improve cross-dataset AFib detection.

Second, the present experiments were conducted retrospectively using publicly available ECG databases and reference beat annotations. Consequently, the reported performance reflects the classification framework under reliable R-peak localization and does not directly quantify the effect of errors introduced by automatic beat detection in noisy ambulatory recordings. Future studies should therefore evaluate the complete pipeline using automatically detected R-peaks and ECG signals acquired under realistic wearable monitoring conditions.

Table 7
Performance Comparison of the proposed MOE-ECG model and state-of-the-art methods for AFib detection on test results. The highest values are highlighted in bold. (LSTM: Long Short-Term Memory, BiLSTM: Bidirectional LSTM, RNN: Recurrent Neural Network, CNN: Convolutional Neural Network, AdaBoost: Adaptive Boosting, LM: Learning Model, CV: Cross Validation, –: Not reported)

Reference	Segment length (heartbeats/seconds)	Model	Train data	Test data	Test performance		
					Acc	Se	F_1-score
Pereira et al. [66]	10 s	LSTM	AFDB	LTAFDB	0.827	0.876	–
Garcia-Isla et al. [67]	60 s	Poincaré Image	–	LTAFDB	0.898	0.946	–
Seo et al. [68]	10 s	ResNet-50	AFDB	LTAFDB	0.849	0.806	0.842
Huang et al. [69]	5 s	Residual Block + Transformer	AFDB	LTAFDB	0.913	0.912	**0.915**
Wen et al. [70]	100 RRIs	Sobolev test statistics + End-to-end LM	LTAFDB	LTAFDB (CV)	0.898	0.959	–
Luo et al. [71]	32 RRIs	LSTM	CPSC2021	LTAFDB	**0.909**	0.912	–
Hu et al. [72]	3 heartbeats	CNN + Transformer	CPSC2021	LTAFDB	0.862	0.657	0.740
Jahan et al. [12]	60 RRIs	AdaBoost	AFDB	LTAFDB	0.903	0.851	0.854
MOE-ECG (proposed)	60 RRIs	MOPSO based ensemble	AFDB	LTAFDB	0.847	**0.982**	0.815

Third, the three databases differ in recording conditions, lead configurations, sampling frequencies, and rhythm-annotation conventions. Although the separation of AFDB, SHDB-AF, and LTAFDB into training, validation, and independent testing provides a useful assessment of cross-dataset generalization, these differences may still influence the observed performance. Evaluation on additional independent cohorts and prospectively collected ECG data would provide stronger evidence of generalizability.

In addition, the use of 50% overlapping windows introduces dependence between neighboring segments from the same recording. While the reported generalization results are obtained on an independent database, future work could investigate recording-level or patient-level evaluation strategies to further characterize performance under stricter independence assumptions.

Finally, computational profiling and prospective device-level evaluation were beyond the scope of this study. Accordingly, wearable and real-time use should be regarded as potential applications rather than demonstrated deployment capabilities. Future work should assess inference time, memory requirements, energy consumption, and end-to-end performance on representative wearable or edge-computing hardware.

4. Conclusion

This study presented MOE-ECG for AFib detection from short ECG segments. The framework uses multi-objective optimization to select a compact subset from 22 candidate classifiers while balancing predictive performance and classifier diversity. The selected outputs are then combined using Dempster–Shafer evidence fusion.

AFDB, SHDB-AF, and LTAFDB were used for training, validation, and independent testing, respectively. Across the three examined segment lengths, the 60-beat configuration provided the most balanced trade-off between temporal context and segment duration and was therefore used as the primary operating point. On LTAFDB, the five-run mean accuracy, sensitivity, and F_1-score were 0.847, 0.982, and 0.815, respectively. The repeated-run baseline comparison also showed higher mean accuracy, precision, and F_1-score for MOE-ECG than for the single-best and unoptimized average ensembles, while these differences are interpreted descriptively because only five independent runs were available.

Across published studies, MOE-ECG showed competitive performance on the independent LTAFDB test set, particularly in sensitivity. Because study designs differ, these results do not establish uniformly superior performance. Overall, the findings indicate that diversity-aware ensemble selection provides a competitive, sensitivity-oriented approach to cross-dataset AFib detection; the main limitations and corresponding directions for future investigation are summarized in Section 3.5.

CRediT authorship contribution statement

Abdolrahman Peimankar: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Naser Hossein Motlagh:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Smith K. Khare:** Writing – review & editing, Visualization, Validation, Methodology. **Nicolai Spicher:** Writing – review & editing, Validation, Methodology. **Helena Dominguez:** Writing – review & editing, Validation, Conceptualization. **Vahid Abolghasemi:** Writing – review & editing, Validation, Methodology. **Koichi Fujiwara:** Writing – review & editing, Validation, Conceptualization. **Daniel Teichmann:** Writing – review & editing, Validation, Resources. **Rahim Rahmani:** Writing – review & editing, Validation, Methodology. **Sadasivan Puthusserypady:** Writing – review & editing, Validation, Resources, Methodology, Conceptualization.

Funding

This research received no external funding

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this work are available on PhysioNet, MIT-BIH Atrial Fibrillation Database, Saitama Heart Database Atrial Fibrillation, and Long-Term Atrial Fibrillation Database.

References

[1] Chugh SS, Havmoeller R, Narayanan K, Singh D, Rienstra M, Benjamin EJ, Gillum RF, Kim Y-H, McAnulty Jr. JH, Zheng Z-J, et al. Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. Circulation 2014;129(8):837–47.

[2] Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation. Chest 2010;137(2):263–72.

[3] Krijthe BP, Kunst A, Benjamin EJ, Lip GYH, Franco OH, Hofman A, Witteman JCM, Stricker BHC, Heeringa J. Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060. Eur Heart J 2013;34(35):2746–51.

[4] Goldberger AL, Goldberger ZD, Shvilkin A. Goldberger's clinical electrocardiography-e-book: A simplified approach. Elsevier Health Sciences; 2023.

[5] De Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng 2004;51(7):1196–206.

[6] Le Heuzey J-Y, Paziaud O, Piot O, Ait Said M, Copie X, Lavergne T, Guize L. Cost of care distribution in atrial fibrillation patients: the COCAF study. Am Heart J 2004;147(1):121–6.

[7] Van Gelder IC, et al. 2024 ESC guidelines for the management of atrial fibrillation developed in collaboration with the European association for cardio-thoracic surgery (EACTS). Eur Heart J 2024;45(36):3314–414.

[8] Lal B, Ajbani M, Sodhro AH, Gravina R, O'Keeffe D. Efficient ECG compression and reconstruction via deep compressive sensing and hardware-efficient models. Array 2026;30:100791.

[9] Acharya UR, Fujita H, Lih OS, Hagiwara Y, Tan JH, Adam M. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inform Sci 2017;405:81–90.

[10] Katsaouni N, Aul F, Krischker L, Schmalhofer S, Hedrich L, Schulz MH. Energy efficient convolutional neural networks for arrhythmia detection. Array 2022;13:100127.

[11] Faiz MMU, Kale I. Removal of multiple artifacts from ECG signal using cascaded multistage adaptive noise cancellers. Array 2022;14:100133.

[12] Jahan MS, Mansourvar M, Puthusserypady S, Wiil UK, Peimankar A. Short-term atrial fibrillation detection using electrocardiograms: a comparison of machine learning approaches. Int J Med Inform 2022;163:104790.

[13] Andersen RS, Peimankar A, Puthusserypady S. A deep learning approach for real-time detection of atrial fibrillation. Expert Syst Appl 2019;115:465–73.

[14] Peimankar A, Puthusserypady S. DENS-ECG: a deep learning approach for ECG signal delineation. Expert Syst Appl 2021;165:113911.

[15] Peimankar A, Jajroodi MJ, Puthusserypady S. Automatic detection of cardiac arrhythmias using ensemble learning. In: TENCON 2019-2019 IEEE region 10 conference. IEEE; 2019, p. 383–8.

[16] Peimankar A, Puthusserypady S. Ensemble learning for detection of short episodes of atrial fibrillation. In: 2018 26th European signal processing conference. IEEE; 2018, p. 66–70.

[17] Prabhakararao E, Dandapat S. Multi-scale convolutional neural network ensemble for multi-class arrhythmia classification. IEEE J Biomed Health Inform 2022;26(8):3802–12.

[18] Sowan B, Zhang L, Houssein EH, Qattous H, Azzeh M, Massad B. DOVE-FELM: a fusion-optimized feature selection and heterogeneous ensemble learning framework for early prediction of chronic kidney disease risk. Array 2025;28:100613.

[19] Kuncheva LI. Combining pattern classifiers: Methods and algorithms. John Wiley & Sons; 2014.

[20] Peimankar A, Puthusserypady S. An ensemble of deep recurrent neural networks for p-wave detection in electrocardiogram. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing. IEEE; 2019, p. 1284–8.

[21] Krawczyk B, Minku LL, Gama J, Stefanowski J, Woniak M. Ensemble learning for data stream analysis: a survey. Inf Fusion 2017;37:132–56.

[22] Coello Coello CA, Pulido GT, Lechuga MS. Handling multiple objectives with particle swarm optimization. IEEE Trans Evol Comput 2004;8(3):256–79.

[23] Habib M, Aljarah I, Faris H, Mirjalili S. Multi-objective particle swarm optimization: theory, literature review, and application in feature selection for medical diagnosis. In: Evolutionary Machine Learning Techniques: Algorithms and Applications. Springer; 2020, p. 175–201.

[24] Rojas MG, Olivera AC, Vidal PJ. Optimising multilayer perceptron weights and biases through a cellular genetic algorithm for medical data classification. Array 2022;14:100173.

[25] Meng J, Huo X, Zhao H, Zhang G, Zhang L, Wang X, Lin J, Zhou S. Multi-modal biological feature selection for parkinson's disease staging based on binary PSO with broad learning. Biomed Signal Process Control 2024;94:106234.

[26] Nguyen BH, Xue B, Zhang M. A survey on swarm intelligence approaches to feature selection in data mining. Swarm Evol. Comput 2020;54:100663.

[27] Singh M, Verma A, Sharma N. Multi-objective noise estimator for the applications of de-noising and segmentation of MRI data. Biomed Signal Process Control 2018;46:249–59.

[28] Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M. Ensemble learning for data stream analysis: A survey. Inf Fusion 2017;37:132–56.

[29] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–20.

[30] Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. In: Computers in cardiology. 10, 1983, p. 227–30.

[31] Tsutsui K, Brimer SB, Ben-Moshe N, Sellal JM, Oster J, Mori H, Ikeda Y, Arai T, Nakano S, Kato R, et al. SHDB-AF: A Japanese Holter ECG database of atrial fibrillation. Sci Data 2025;12(1):454.

[32] Tsutsui K, Biton Brimer S, Behar J. SHDB-AF: a japanese holter ECG database of atrial fibrillation. PhysioNet 2025. Version 1.0.1.

[33] Petrutiu S, Sahakian AV, Swiryn S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace 2007;9(7):466–70.

[34] Moody G, Pollard T, Moody B. WFDB software package. PhysioNet 2022. Version 10.7.0.

[35] Clifford GD, Azuaje F, McSharry PE, editors. Advanced methods and tools for ECG data analysis. Artech House; 2006.

[36] Gierałtowski J, Żebrowski J, Baranowski R. Multiscale multifractal analysis of heart rate variability recordings with a large number of occurrences of arrhythmia. Phys Rev E-Stat Nonlinear Soft Matter Phys 2012;85(2):021915.

[37] Faust O, Shenfield A, Kareem M, Tan RS, Fujita H, Acharya UR. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. Comput Biol Med 2018;102:327–35.

[38] Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, Schwartz PJ. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Eur Heart J 1996;17(3):354–81.

[39] Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258.

[40] Malik M, Camm AJ. Heart rate variability and clinical cardiology. Br Heart J 1994;71(1):3–6.

[41] Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Am Stat Assoc 1993;88(424):1273–83.

[42] Voss A, Schulz S, Schroeder R, Baumert M, Caminal P. Methods derived from nonlinear dynamics for analysing heart rate variability. Phil Trans R Soc A 2009;367:277–96.

[43] Grégoire J-M, Gilon C, Marelli F, Bersini H, Groben L, Nguyen T, Deruyter B, Godart P, Carlier S. Short-term atrial fibrillation onset prediction using machine learning. Eur Heart J-Digital Health 2025;6(6):1159–68.

[44] Brennan M, Palaniswami M, Kamen P. Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability?. IEEE Trans Biomed Eng 2001;48(11):1342–7.

[45] Langley P, Dewhurst M, Di Marco LY, Adams P, Dewhurst F, Mwita JC, Walker R, Murray A. Accuracy of algorithms for detection of atrial fibrillation from short duration beat interval recordings. Med Eng Phys 2012;34(10):1441–7.

[46] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018, arXiv preprint arXiv:1802.03426.

[47] Maaten Lvd, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(Nov):2579–605.

[48] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 2003;51(2):181–207.

[49] Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Lecture Notes in Computer Science, 1857, Springer; 2000, p. 1–15.

[50] Opitz D, Maclin R. Popular ensemble methods: An empirical study. J Artificial Intelligence Res 1999;11:169–98.

[51] Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence 2021;35(8):6679–87.

[52] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research 2012;13:281–305.

[53] Claesen M, De Moor B. Hyperparameter search in machine learning. 2015, arXiv preprint arXiv:1502.02127.

[54] Kuncheva LI. A theoretical study on six classifier fusion strategies. IEEE Trans Pattern Anal Mach Intell 2002;24(2):281–6.

[55] Deb K. Multi-objective optimisation using evolutionary algorithms: An introduction. In: Multi-objective evolutionary optimisation for product design and manufacturing. Springer; 2011, p. 3–34.

[56] Coello Coello CA, Lechuga MS. MOPSO: a proposal for multiple objective particle swarm optimization. In: Proceedings of the 2002 congress on evolutionary computation. vol. 2, IEEE; 2002, p. 1051–6.

[57] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans Evol Comput 1999;3(4):257–71.

[58] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of iCNN'95 - international conference on neural networks, vol. 4. 1995, p. 1942–8.

[59] Eberhart RC, Shi Y. Particle swarm optimization: developments, applications and resources. In: Proceedings of the 2001 congress on evolutionary computation. vol. 1, IEEE; 2001, p. 81–6.

[60] Engelbrecht AP. Computational intelligence: An introduction. 2. Wiley; 2007.

[61] Knowles JD, Corne DW. Approximating the nondominated front using the Pareto archived evolution strategy. Evol Comput 2000;8(2):149–72.

[62] Shafer G. A mathematical theory of evidence. Princeton: Princeton University Press; 1976.

[63] Massey Jr. FJ. The Kolmogorov-Smirnov test for goodness of fit. J Amer Statist Assoc 1951;46(253):68–78.

[64] Dekking FM, Kraaikamp C, Lopuhaä HP, Meester LE. A modern introduction to probability and statistics: Understanding why and how. London: Springer; 2005.

[65] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.

[66] Pereira R, Andreão RV. Inter-patient detection of atrial fibrillation in short ECG segments based on LSTM network with multiple input layers. Res Biomed Eng 2022;38(2):465–76.

[67] Garcia-Isla G, Corino V, Mainardi L. Poincaré plot image and rhythm-specific atlas for atrial bigeminy and atrial fibrillation detection. IEEE J Biomed Health Inform 2021;25(4):1093–100.

[68] Seo H-C, Oh S, Kim H, Joo S. ECG data dependency for atrial fibrillation detection based on residual networks. Sci Rep 2021;11(1):18256.

[69] Huang Z, Yang S, Zou Q, Gao X, Chen B. A portable household detection system based on the combination of bidirectional LSTM and residual block for automatical arrhythmia detection. Biomed Eng/Biomed Tech 2024;69(2):167–79.
[70] Wen H, Kang J. A comparative study on neural networks for paroxysmal atrial fibrillation events detection from electrocardiography. J Electrocardiol 2022;75:19–27.
[71] Luo C, Zhang K, Hu Y, Li X, Cao S, Jin Y, Ren P, Rao N. Directional statistics-inspired end-to-end atrial fibrillation detection model based on ECG rhythm. Expert Syst Appl 2024;247:123112.
[72] Hu Y, Feng T, Wang M, Liu C, Tang H. Detection of paroxysmal atrial fibrillation from dynamic ECG recordings based on a deep learning model. J Pers Med 2023;13(5):820.