Volume 8 Nomor 2 Juli 2026

JJEEE

e-ISSN : 2715-0887
p-ISSN : 2654-7813

Jambura Journal of Electrical and Electronics Engineering

Perbandingan Seleksi Fitur Filter dan SHAP untuk Klasifikasi Atrial Fibrillation berbasis ECG

Comparison of Filter and SHAP Feature Selection for ECG-based Atrial Fibrillation Classification

Novie Theresia Br. Pasaribu
Program Studi Teknik Elektro
Universitas Kristen Maranatha
Bandung, Indonesia
novie.theresia@eng.maranatha.edu

Elizabeth Fabiola Wijaya
Program Studi Teknik Elektro
Universitas Kristen Maranatha
Bandung, Indonesia
2122005@eng.maranatha.edu

Che Wei Lin
Department of Biomedical Engineering
National Cheng Kung University
Tainan, Taiwan
lincw@mail.ncku.edu.tw

Diterima : Juni 2026
Disetujui : Juli 2026
Dipublikasi : Juli 2026

Abstrak*—Atrial Fibrillation (AF) merupakan jenis aritmia yang kasusnya terus meningkat secara global dan berpotensi menimbulkan komplikasi serius seperti stroke dan serangan jantung. Deteksi dini berbasis sinyal ECG secara otomatis menjadi sangat penting. Penelitian ini membandingkan tiga metode seleksi fitur, yaitu Pearson Correlation (PC), Mutual Information (MI), dan Shapley Additive Explanations (SHAP), untuk klasifikasi AF pada sinyal ECG 2-lead menggunakan XGBoost. Dataset yang digunakan adalah MIT-BIH Atrial Fibrillation Database dengan 23 rekaman ECG, menghasilkan 34.312 segmen data berdurasi 10 detik. Preprocessing menggunakan Stationary Wavelet Transform (SWT) dan normalisasi Min-Max. Terdapat 50 ekstraksi fitur yang dihasilkan. Setiap metode diuji pada 4 ukuran subset fitur (5, 10, 15, dan 20 fitur). Model dioptimasi menggunakan GridSearchCV dengan 5-fold Stratified Cross-Validation. Hasil menunjukkan bahwa seleksi fitur SHAP mengungguli PC dan MI pada subset 10, 15, dan 20. SHAP dengan 20 fitur mencapai performa tertinggi (akurasi 97,87%; F1-score 96,78%; ROC-AUC 0,9971), sedangkan SHAP dengan 15 fitur menawarkan trade-off performa–dimensionality terbaik: akurasi 97,84%, presisi 97,85%, recall 95,62%, F1-score 96,72%, dan ROC-AUC 0,9965, dengan reduksi dimensi 70% (0,03% di bawah SHAP-20 pada akurasi, dengan presisi lebih tinggi). Keunggulan SHAP terhadap PC dan MI terbukti signifikan secara statistik (uji McNemar dan DeLong, p < 0,05) pada subset 15 dan 20 fitur; SHAP dengan 20 fitur mencapai ROC-AUC yang tidak berbeda secara signifikan dari baseline 50 fitur (uji DeLong, p = 0,4997).

***Kata Kunci*—Atrial Fibrillation; Feature Selection; SHAP; ECG Classification; XGBoost**

Abstract*—Atrial Fibrillation (AF) is a type of arrhythmia whose prevalence continues to rise globally and can lead to serious complications such as stroke and heart attack. Early detection based on ECG signals is therefore crucial. This study compares three feature selection methods, namely Pearson Correlation (PC), Mutual Information (MI), and Shapley Additive Explanations (SHAP), for classifying AF from 2-lead ECG signals using XGBoost. The dataset used was the MIT-BIH Atrial Fibrillation Database, comprising 23 ECG recordings, yielding 34,312 10-second data segments. Preprocessing used Stationary Wavelet Transform (SWT) and Min-Max normalisation. A total of 50 features were extracted. Each method was tested on 4 feature subset sizes (5, 10, 15, and 20 features). The model was optimised using GridSearchCV with 5-fold Stratified Cross-Validation. Results showed that SHAP outperformed PC and MI in subsets of 10, 15, and 20 features. SHAP with 20 features achieved the highest performance (97.87% accuracy; F1 score 96.78%; ROC-AUC 0.9971), while the 15-feature SHAP offered the best performance–dimensionality trade-offs: 97.84% accuracy, 97.85% precision, 95.62% recall, 96.72% F1-score, and 0.9965 ROC-AUC, with a 70% dimensional reduction (0.03% below SHAP-20 on accuracy, with higher precision). SHAP's superiority over PC and MI was statistically significant (McNemar and DeLong tests, p < 0.05) in subsets of 15 and 20 features; SHAP with 20 features achieved a ROC-AUC that did not differ significantly from the baseline of 50 features (DeLong test, p = 0.4997).

***Keywords*—Atrial Fibrillation; Feature Selection; SHAP; ECG Classification; XGBoost**

I. INTRODUCTION

People with Atrial Fibrillation (AF) and Atrial Flutter (AFL) have more than doubled from 22.21 million cases in 1990 to 52.55 million in 2021 [1]. This trend is projected to continue to increase until 2035, with an estimated increase in incidence of 4.07% [1]. Atrial Fibrillation (AF) is a type of arrhythmia characterised by impaired heart rhythm activity due to decreased atrial function or very irregular atrial depolarisation [2], [3]. AF often does not cause symptoms and can lead to serious complications, such as stroke and heart attack [4]. Therefore, early detection of AF is very important to prevent more severe clinical risks.

AF can be detected from electrocardiogram (ECG) signals by analysing the PQRST complexes. AF is usually characterised by the loss of P waves, the appearance of fairly rapid oscillations, and irregular RR intervals [2]. However, manual AF detection is challenging and prone to

misdiagnosis because the ECG signal is weak and easily affected by noise [5], [6]. Thus, an automated approach capable of extracting key characteristics from ECG signals is needed to improve the accuracy of AF detection.

Machine Learning (ML) approaches have shown great potential for automating the classification of Electrocardiogram (ECG) signals to detect AF [7]. Various ML algorithms, including ensemble learning methods such as Random Forest, LightGBM, and Gradient Boosting, have been applied successfully [8]. Bhatt et al. show an ML approach to predict heart disease based on clinical indicators [9]. For ECG-based AF detection in particular, Sharmin et al. used XGBoost with a hand-crafted set of interpretable features and achieved an accuracy of about 96% with an F1-score of 0.83 for the AF class, equivalent to the best classifier of the 2017 PhysioNet Challenge despite using only 48 features compared to 150–622 features [4].

The process of extracting features from ECG signals often yields many features that are not entirely relevant to classification performance. Irrelevant features can degrade model performance through the curse of dimensionality and increase the risk of overfitting [10]. Previous research has shown that proper feature selection can improve classification accuracy while significantly reducing computational complexity [11]. Feature selection aims to select the most informative subset of features, resulting in more accurate, more computationally efficient models.

In the ML literature, feature selection methods are generally categorised into three paradigms: filter, wrapper, and embedded [12]. Filter methods, such as Pearson Correlation, evaluate features using statistical measures relative to target variables without involving a classification model, so they are quick but do not consider interactions between features [10], [12]. Gong et al. showed that combining Pearson Correlation and normalised Mutual Information can improve the effectiveness of feature selection on classification tasks [14]. This merger is motivated by the limitations of using a single indicator: Pearson Correlation only captures linear dependencies, while Mutual Information captures non-linear dependencies but is sensitive to density estimation in continuous data [15]. Shapley Additive Explanations (SHAP) is a post-hoc, model-agnostic interpretability method that evaluates the contribution of each feature to model prediction based on Shapley's concept of value [16]. Because its attribution reflects the model's behaviour, SHAP captures the influence of features, including the effects of inter-feature interactions, so SHAP has been widely applied [17].

Wang et al. compared SHAP-based and importance-based feature selection for fraud detection across different feature subset sizes, but this comparison was limited to tabular fraud data and did not include a filter paradigm [13]. Cao et al. used multiple feature selection methods by combining Pearson correlation and feature importance ranking for cardiovascular predictions, but the two were fused in a single pipeline without head-to-head comparisons between paradigms [11]. Therefore, in this study, a systematic comparative analysis was carried out on three feature selection methods that include two different paradigms, namely Pearson Correlation (linear filter), Mutual Information (non-linear filter), and SHAP (model-based/post-hoc), for the classification of AF in 2-lead ECG signals. All three feature selection methods were tested on four feature subset sizes (5, 10, 15, and 20 features out of a total of 50 baseline features) using XGBoost.

The contributions to this study are: (i) A comparison between two feature selection paradigms, namely: filter model (PC and MI) versus post-hoc model (SHAP), for AF classification on 2-lead ECG signals; (ii) Evaluation of features by statistical significance test (McNemar and DeLong); and (iii) Analysis of confusion matrix and False Positive/False Negative.

II. METHOD

The research method in this study consists of several stages: a literature review (on AF, ML, feature extraction, feature selection, and XGBoost), AF classification design, analysis of results, and, finally, conclusions. The stages of the AF classification design process are shown in Figure 1.

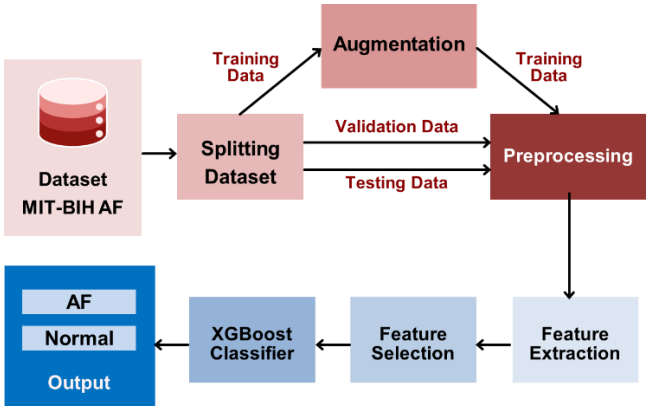


Figure 1. Block Diagram of AF Classification Diagram Using XGBoost

In the initial stage, the MIT-BIH Atrial Fibrillation dataset is split into three parts: training, validation, and test. To address class imbalances, synthetic data for minority classes is generated using the Synthetic Minority Over-Sampling Technique (SMOTE). Then, all data (training, validation, and test) undergo preprocessing, followed by feature extraction and selection. Based on the feature selection results, the selected features are used as input to the XGBoost Classifier for classification. The system's final output is a two-class classification: Normal (N) and Atrial Fibrillation (AF).

A. Dataset

The dataset used in this study is the MIT-BIH Atrial Fibrillation Database, which contains 25 two-lead ECG recordings sampled at 250 Hz. In this study, 23 of the 25 recordings were used, with a focus on two classes: Atrial Fibrillation (AF) and Normal (N). Each ECG recording in the time windowing is 10 seconds long with no overlap, resulting in a total of 34,312 data points, comprising the AF and Normal (N) classes. The dataset is divided into three parts: training, validation, and test, with a 60:20:20 split. The overall data distribution is shown in Table 1.

TABLE 1. TABLE OF DATA DISTRIBUTION

Class	Data		
	Training	*Validation*	*Testing*
N	13,736	4,580	4,580
AF	6,850	2,283	2,283
Total Data	**20,586**	**6,863**	**6,863**

Training data is 20,586, validation data is 6,863, and test data is 6,863.

Data augmentation using the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training data to address class imbalances and the model's tendency to overfit majority classes. After data augmentation with SMOTE, the training data increased to 27,472, consisting of 13,736 N classes and 13,736 AF classes.

B. *Preprocessing Process*

In this study, preprocessing consists of two main stages: Stationary Wavelet Transform (SWT) for noise reduction and data normalisation. SWT is used to filter noise in non-stationary ECG signals (Figure 2). The SWT parameters used in this study are Daubechies wavelet 4 (db4), level 3 decomposition, and soft-thresholding with a threshold value of 0.04, to preserve the characteristics of the ECG signal while eliminating noise [18], [19]. Normalisation is a preprocessing step that transforms the data into a common range, so that features with larger values do not dominate those with smaller values [20]. Normalisation is performed by mapping the ECG signal's amplitude range to [0, 1] using Min-Max normalisation.

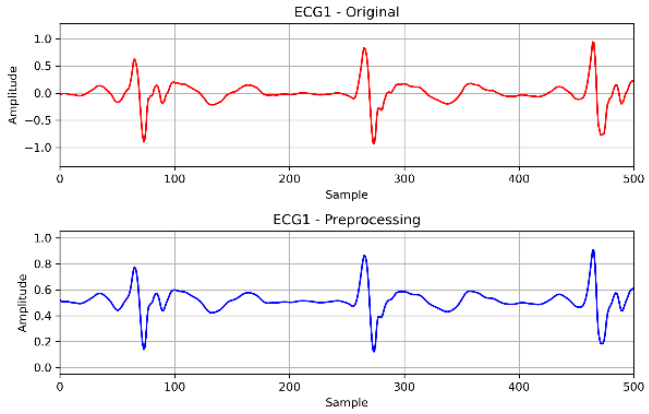


Figure 2. Preprocessing Process of ECG Signal Result

C. *Feature Extraction*

Feature extraction is a critical stage in the ECG signal classification system because it aims to extract key characteristics of the signal that effectively represent the heart's physiological condition. At this stage, three feature extraction methods are applied: Morphological ECG, Statistical Characteristics, and Wavelet Decomposition. For extracting ECG Morphology and Statistical features, the focus is on detecting R-peaks and R-R intervals in ECG signals.

In this study, to detect R-peaks in the ECG signal, we propose using the average amplitude of the ECG signal and a dynamically adjusted threshold based on its characteristics. This method is referred to hereinafter as the Adaptive Thresholding (AT) method. The following are the stages of the process for detecting R-peaks in the ECG signal with AT:

- The initial step is to calculate the average amplitude of the ECG signal (mean_val) and the maximum amplitude of the ECG signal (max_val).
- Then calculate the value of the signal range, which is the difference between max_val and mean_val (see formula (1)).

$$range = \max\_val - mean\_val \quad (1)$$

- Calculate the signal offset value to determine the adaptive threshold for the signal range (see formula (2)). The scale value ranges from 0 to 1.
  In this study, a scale of 0.4 was used, meaning the offset is 40% of the range.

$$offset = scale \times range \quad (2)$$

- Finally, calculate the threshold value: the threshold used to determine the R-peak in the ECG signal; see equation (3). If the value exceeds the threshold, it is considered an R-peak in the ECG signal.

$$threshold = mean\_val + offset \quad (3)$$

The AT method is proposed because it is computationally simple, and the threshold is dynamically adjusted based on the amplitude of the observed ECG sample, without bandwidth filtering or cascaded integration as in the Pan-Tompkins algorithm [21]. Quantitative comparisons of R-peak detection between the AT method and other methods, such as Pan-Tompkins [21]/Hamilton [22], are not discussed in this study (as a next direction for the study) because the focus is on feature selection.

An illustration of the AT method for detecting the R-peak of the ECG signal is shown in Figure 3. After obtaining the max_val of the signal (green line) as 0.9, the mean_val (red line) is 0.1, so the range = 0.9 - 0.1 = 0.8. If scale = 0.4, then offset = 0.4 × 0.8 = 0.32. Finally, a threshold (orange line) of 0.1 + 0.32 = 0.42 is obtained.

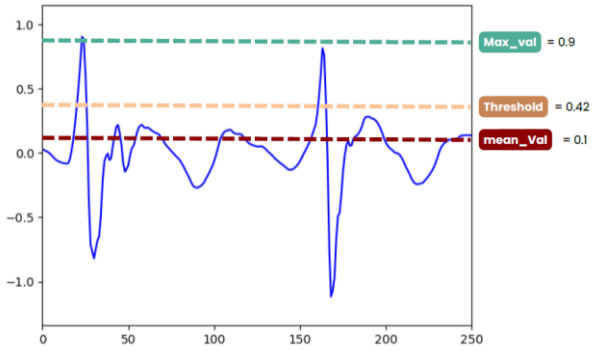


Figure 3. Illustration of R-Peak Detection using the AT method

Here are ECG signal results showing the successfully detected R-peak (see Figure 4).

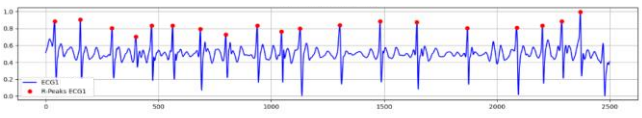


Figure 4. Example of the Results of R-peak Detection

1) Morphological ECG Feature Extraction

Extracting this feature yields 2 features: the number of R-Peaks detected in the ECG1 lead (R_Peak_ECG1) and the number of R-Peaks detected in the ECG2 lead (R_Peak_ECG2).

2) Statistical Characteristics Feature Extraction

Statistical characteristics were chosen because they provide a concise, informative representation of ECG signal dynamics. Statistical characteristics are calculated for each ECG signal window, following the formulation of Mengting Shen et al. [23]. In this study, the Statistical Characteristics feature was extracted by calculating statistical parameters of the ECG signal and the RR interval (the distance between successive R peaks).

The features are counted as many as 16 features, consisting of:

- Mean amplitude on ECG1 and ECG2 signal (mean_ECG1, mean_ECG2).
- Root Mean Square (RMS) amplitude on ECG1 and ECG2 signal (rms_ECG1, rms_ECG2).
- Mean of interval RR on ECG1 and ECG2 signals (RR_mean_ECG1, RR_mean_ECG2).
- Standard Deviation (Std) of RR interval on ECG1 and ECG2 signal (RR_std_ECG1, RR_std_ECG2).
- Skewness distribution of RR interval on signal ECG1 and ECG2 (RR_skew_ECG1, RR_skew_ECG2).
- Distribution Kurtosis of RR interval on ECG1 and ECG2 signal (RR_kurt_ECG1, RR_kurt_ECG2).
- Mean delta of RR interval on ECG1 and ECG2 signal (Delta_RR_mean_ECG1, Delta_RR_mean_ECG2).
- Standard deviation of delta RR interval on ECG1 and ECG2 signal (Delta_RR_std_ECG1, Delta_RR_std_ECG2).

3) Wavelet Decomposition Feature Extraction

Wavelet decomposition for feature extraction enables simultaneous time- and frequency-domain analysis (time–frequency analysis), making it suitable for capturing the characteristics of non-stationary ECG signals. The ECG signal was decomposed using Daubechies order-2 (db2) wavelets to level 2. This decomposition follows the approach of Mandala et al. (2023) [19]. The decomposition process is carried out in stages: at level 1, the original signal is decomposed into Approximation Coefficient (A1) and Detail Coefficient (D1); at level 2, Coefficients A1 and D1 are further decomposed, resulting in A2 and D2 on each branch. The level 2 decomposition yielded 4 wavelet sub-bands for each ECG lead: A1, A2, D1, and D2.

Each subband decomposition result is analysed using statistical parameters to characterise the distribution of wavelet coefficients in each subband. The extracted parameters are Minimum (Min), Maximum (Max), Mean (Mean), and Standard Deviation (Std), yielding 32 features for each ECG signal:

- Mean A1 on ECG1 and ECG2 signals (ECG1_Mean_A1, ECG2_Mean_A1).
- Min A1 on ECG1 and ECG2 signals (ECG1_Min_A1, ECG2_Min_A1)
- Max A1 on ECG1 and ECG2 signals (ECG1_Max_A1, ECG2_Max_A1).
- Std A1 on ECG1 and ECG2 signals (ECG1_Std_A1, ECG2_Std_A1).
- Mean A2 on ECG1 and ECG2 signals (ECG1_Mean_A2, ECG2_Mean_A2).
- Min A2 on ECG1 and ECG2 signals (ECG1_Min_A2, ECG2_Min_A2).
- Max A2 on ECG1 and ECG2 signals (ECG1_Max_A2, ECG2_Max_A2).
- Std A2 on ECG1 and ECG2 signals (ECG1_Std_A2, ECG2_Std_A2).
- Mean D1 on ECG1 and ECG2 signals (ECG1_Mean_D1, ECG2_Mean_D1).
- Min D1 on ECG1 and ECG2 signals (ECG1_Min_D1, ECG2_Min_D1).
- Max D1 on ECG1 and ECG2 signals (ECG1_Max_D1, ECG2_Max_D1).
- Std D1 on ECG1 and ECG2 signals (ECG1_Std_D1, ECG2_Std_D1).
- Mean D2 on ECG1 and ECG2 signals (ECG1_Mean_D2, ECG2_Mean_D2).
- Min D2 on ECG1 and ECG2 signals (ECG1_Min_D2, ECG2_Min_D2).
- Max D2 on ECG1 and ECG2 signals (ECG1_Max_D2, ECG2_Max_D2).
- Std D2 on ECG1 and ECG2 signals (ECG1_Std_D2, ECG2_Std_D2).

The total feature extraction comprises 50 features, consisting of 2 Morphological ECG feature extractions, 16 Statistical Characteristics feature extractions, and 32 Wavelet Decomposition feature extractions.

D. Feature Selections

In this study, three feature selection methods were used: Pearson Correlation (PC), Mutual Information (MI), and Shapley Additive Explanations (SHAP).

1) Pearson Correlation Feature Selection

Pearson Correlation (PC)-based feature selection measures the strength of the linear relationship between a feature and a target variable. As a filter method, the PC works independently of the classifier (model-free), so that the computation is very efficient [10].

2) Mutual Information Feature Selection

Mutual Information (MI)-based feature selection measures the amount of information shared between a feature and the target variable [24], and is capable of capturing dependencies both linear and non-linear [15]. However, MI estimates are sensitive to density estimates in high-dimensional continuous data [25].

3) SHAP Feature Selection

The Shapley Additive Explanations (SHAP) method calculates each feature's marginal contribution to the model's prediction using the Shapley value concept from

cooperative game theory [16]. For tree-based models such as XGBoost, TreeExplainer is used to efficiently compute SHAP values [26]. The mean absolute SHAP value per feature is used as the global importance score [26]. SHAP is post hoc and model-based: it is calculated after the model is fully trained and thus leverages the model's internal structure and feature interactions [17].

For reproducibility, the entire process (MI, XGBoost, and cross-validation fold estimation) uses a random seed = 42.

The three feature selection methods were each tested with four feature subset (N) sizes, namely 5, 10, 15, and 20 features, selected from a total of 50 existing feature extracts (a reduction in the number of features by 90%, 80%, 70%, and 60% of the total features), plus one baseline using all features without selection. There were 13 trials in total (Table 2).

TABLE 2. TABLE OF EXPERIMENTS TOTAL

# Features	Feature Selection		
	PC	*MI*	*SHAP*
5	v	v	v
10	v	v	v
15	v	v	v
20	v	v	v
50	Baseline (all features)		

During training, feature rankings are computed, and the top N features (N=5, 10, 15, 20) are selected. Next, the model is evaluated on the test data, and its performance metrics are computed.

E. XGBoost Classifier

XGBoost (eXtreme Gradient Boosting) is an ensemble method for building prediction models by gradually combining multiple decision trees. It trains each new tree to predict the residuals (errors) of the previous combined model, so that each added tree focuses on correcting the shortcomings of the existing model [27].

In this study, a hyperparameter search was conducted using GridSearchCV with 5-fold Stratified Cross-Validation. A total of 108 experimental combinations were evaluated, and the configuration with the highest average cross-validation accuracy was selected as the hyperparameter for the selected model. Grid Search was chosen because it is exhaustive in a finite parameter space, producing deterministic and reproducible results, in line with the study of hyperparameter optimisation in clinical data-driven prediction models [28]. The parameter values in this study are shown in Table 3.

TABLE 3. TABLE OF XGBOOST HYPERPARAMETERS

Parameters	Values	Total
Number of decision trees (n_estimators)	100, 200, 300	3
Depth of max. tree (max_depth)	4, 6, 8	3
Learning rate (learning_rate)	0.05, 0.1, 0.2	3
Proportion of sample per tree (subsample)	0.8, 1.0	2
Proportion of features per tree (colsample_bytree)	0.8, 1.0	2
Total	**3×3×3×2×2**	**108**

F. Evaluation Metrics

Five evaluation metrics were used: Accuracy, Precision, Recall, F1-Score, and ROC-AUC [26]. In the context of AF detection, Recall has high clinical significance because a failure to detect AF (a false negative) is potentially more dangerous than a false alarm.

To assess performance differences between methods, this study used the McNemar test for accuracy and the DeLong test for ROC-AUC. Both are suitable for predictions that are correlated with the same test sample. In addition, the computation time for each configuration is recorded.

III. RESULTS AND DISCUSSION

A. Hyperparameter XGBoost

The XGBoost hyperparameter search was conducted with GridSearchCV using 5-fold Stratified Cross-Validation, yielding 108 parameter combinations. The five best configurations, ranked by average cross-validation accuracy, are shown in Table 4.

TABLE 4. TABLE OF TOP FIVE GRIDSEARCHCV CONFIGURATIONS

No	Parameters					CV Acc (%)
	n est	*Max depth*	*Learning rate*	*Sub sample*	*Colsample_bytree*	
1	300	6	0.2	0.8	1.0	98.89±0.07
2	300	8	0.1	1.0	0.8	98.87±0.09
3	300	8	0.1	0.8	1.0	98.87±0.13
4	300	8	0.1	0.8	0.8	98.86±0.12
5	300	8	0.2	1.0	1.0	98.86±0.14

Among the three configurations (n_estimators = 100, 200, 300), n_estimators = 300 delivers the best performance. As shown in Table 4, the top five configurations all set n_estimators to 300, with a low standard deviation in cross-validation accuracy (0.07%–0.14%), indicating high stability across cross-validation folds. The best first-order configuration uses n_estimators=300; max_depth=6; learning_rate=0.2; subsample=0.8; colsample_bytree=1.0, achieving an average cross-validation accuracy of 98.89%±0.07%. This configuration will be used for the entire subsequent experiment.

B. Results of Feature Selection using XGBoost

After the model's hyperparameters were set, a test was conducted to search for feature selection, with 13 experiments. For detailed results, see Table 5. A graph comparing overall accuracy for each feature-reduction and feature-selection method is shown in Figure 5.

Experiments with only 5 features (a 90% reduction) showed that all methods experienced performance degradation. PC Feature Selection (88.12%) achieved the highest accuracy, followed by SHAP (83.90%) and MI (75.72%). The gap between the best and worst methods was 12.40%.

At 10 features (an 80% reduction), the difference between methods was significant. SHAP achieved the highest accuracy at 96.82% (a +12.92% increase compared with 5 features), while other methods improved more modestly: PC accuracy reached 90.19% (+2.07%), and MI accuracy reached 79.40% (+3.68%).

TABLE 5. TABLE OF FEATURE SELECTION PERFORMANCE RESULTS

Feature Selection	# Features	Performance Metric				
		Acc (%)	*Prec (%)*	*Recall (%)*	*F1-Score (%)*	*AUC*
Baseline	50	**98.32**	**97.59**	**97.37**	**97.48**	**0.9974**
PC	5	88.12	82.33	81.87	82.10	0.9455

Feature Selection	# Features	Performance Metric				
		Acc (%)	*Prec (%)*	*Recall (%)*	*F1-Score (%)*	*AUC*
	10	90.19	93.61	75.69	83.70	0.9599
	15	93.37	93.69	85.85	89.60	0.9808
	20	97.26	96.21	95.53	95.87	0.9955
MI	5	75.72	85.83	32.37	47.01	0.9055
	10	79.40	84.03	47.00	60.28	0.9282
	15	92.80	95.99	81.78	88.32	0.9853
	20	96.26	95.88	92.73	94.28	0.9923
SHAP	5	83.90	87.33	60.36	71.38	0.9392
	10	96.82	95.27	95.18	95.22	0.9947
	15	**97.84**	**97.85**	**95.62**	**96.72**	**0.9965**
	20	**97.87**	**97.59**	**95.97**	**96.78**	**0.9971**

Meanwhile, across 15 features (70% reduction), SHAP achieved the highest accuracy at 97.84%, followed by PC at 93.37% and MI at 92.80%. The SHAP accuracy differs from the baseline of 98.32% by 0.48%. The SHAP precision is 97.85%, slightly higher than the baseline of 97.59%. The ROC-AUC for SHAP 15 is 0.9965, close to the baseline of 0.9974.

At 20 features (60% reduction), SHAP achieves the highest accuracy, and the feature selection method converges. PC is 97.26%, SHAP is 97.87%, and MI is 96.26%.

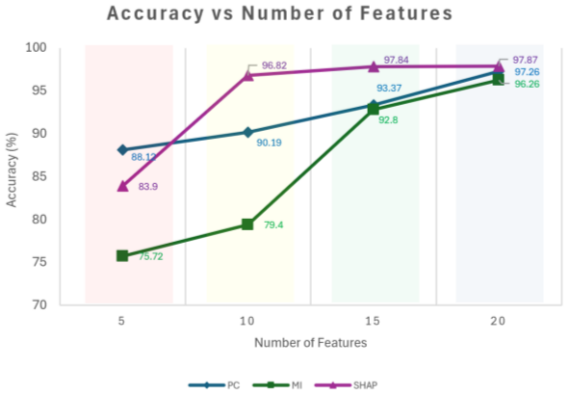


Figure 5. Comparison Graph of the Accuracy of each Feature Selection Method

Figure 5 compares the accuracy of each feature selection method. When a model uses 5 features, PC is the most prominent. However, when the feature count is 10, 15, or 20, SHAP outperforms PC and MI. The most significant result for SHAP compared with the other methods is that, when the feature count is 10, SHAP achieves 96.82% accuracy; PC achieves 90.19%; and MI achieves only 79.40% (a difference of 17.42%). SHAP feature selection with 15 features (97.84%) and 20 features (97.87%) achieved the best accuracy, compared to CS with 20 features (97.26%), MI with 20 features (96.26%), and PC with 20 features (97.26%).

SHAP-20 achieved the highest accuracy of 97.87%, followed by SHAP-15 at 97.84%. An increase from 15 to 20 features adds only 0.03% accuracy, while precision drops (from 97.85% to 97.59%). SHAP-15 provides a 70% reduction (compared to 60% in SHAP-20), while SHAP-20 provides the best performance.

C. Computational Time Analysis

The training times for each configuration are shown in Table 6. Reducing the feature count decreases training time: the baseline 50-feature configuration takes about 1.10 seconds, while the 15-feature configuration takes about 0.54 seconds (a reduction of about 51%). All three methods have almost the same training time for the same number of features.

TABLE 6. TRAINING TIME PER FEATURE SELECTION METHOD

Feature Selection	Number of Features	Training Time (seconds)
Baseline	50	1.10
PC	15	0.54
	20	0.60
MI	15	0.56
	20	0.65
SHAP	15	0.54
	20	0.59

D. Statistical Significance Test

This study uses performance metrics accuracy and ROC-AUC, which produce different analyses. Paired tests on one test set were chosen because the evaluation was carried out on fixed test data, so that only one performance value per model was available. To confirm the difference in performance between methods, two paired significance tests were performed on identical test data (6,863 samples): the McNemar test for accuracy difference [29] and the DeLong test for ROC-AUC difference [30]. Tests were performed between SHAP and both filter methods (PC, MI) as well as between each method and a baseline of 50 features, on both subset sizes (15 and 20). The results of the McNemar test are presented in Table 7 and the results of the DeLong test in Table 8.

TABLE 7. MCNEMAR TEST (ACCURACY)

Number of Features	Model		χ^2	p (sig.)
	A	B		
15	SHAP	PC	227.82	<0.0001
	SHAP	MI	282.05	<0.0001
	SHAP	Baseline	10.56	0.0012
20	SHAP	PC	12.93	0.0003
	SHAP	MI	57.35	<0.0001
	SHAP	Baseline	11.11	0.0009

In Table 7, the results of the McNemar test are shown, which assesses whether the two models produce different patterns of misclassification on the same test set. The χ^2 statistic is calculated from the number of samples that are classified as true by one model but false by another. A large value of χ^2 indicates a difference in accuracy. Against both filter methods, SHAP showed very significant differences in both subset sizes (SHAP vs PC and SHAP vs MI, overall $p < 0.001$), with χ^2 highest in the 15-feature subset ($\chi^2 = 282.05$ for MI). This shows that the SHAP method has the best accuracy. As for the baseline, the difference in SHAP accuracy was also significant, with a much smaller χ^2 ($\chi^2 = 10.56$ in subset 15 and $\chi^2 = 11.11$ in subset 20).

TABLE 8. DELONG TEST (ROC-AUC)

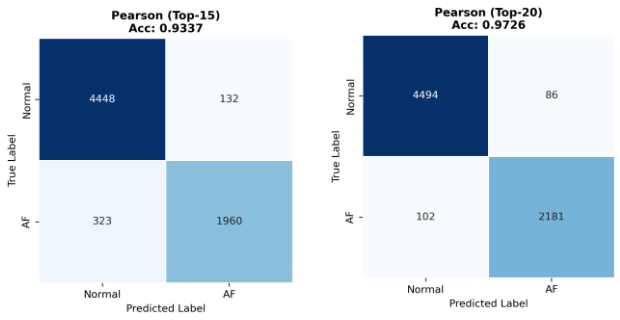
# Ft	Model		AUC		z	p (sig.)
	A	B	A	B		
15	SHAP	PC	0.9965	0.9808	11.70	<0.0001
	SHAP	MI	0.9965	0.9853	11.19	<0.0001
	SHAP	Baseline	0.9965	0.9974	-2.58	0.0100
20	SHAP	PC	0.9971	0.9955	4.34	<0.0001
	SHAP	MI	0.9971	0.9923	7.01	<0.0001
	SHAP	Baseline	0.9971	0.9974	-0.67	**0.4997**

The results of the DeLong test (Table 8) compared the ROC-AUC values on the same test data by considering the correlation between predictions. A z-statistic indicates the direction of the difference (positive means that Model A's AUC is higher), while a p-value determines its significance. Compared to the two filter methods, SHAP had a significantly higher AUC at both subset sizes ($p<0.001$ overall), with the largest difference in the 15-feature subset ($z=11.70$ for PC; $z=11.19$ for MI). Compared to baseline, SHAP-20 obtained $z = -0.67$ with $p = 0.4997$, which means that its AUC did not differ significantly from the 50-feature model even though 60% of features were reduced. Meanwhile, SHAP-15 showed a significant difference ($z=-2.58$; $p=0.0100$), but the absolute AUC difference was only 0.0009 (0.9965 vs 0.9974). Both filter methods still experienced significant AUC degradation compared to baseline in both subset sizes ($p<0.001$).

E. Confusion Matrix & False Positive/ False Negative Analysis

The sixth configuration confusion matrix (three methods with two subset sizes, 15 and 20) of the classification results in the test data (6,863 samples) is seen in Figure 6. Each matrix shows four cells: True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). In the SHAP matrix, diagonal dominance (TN and TP) is seen, and the FN value is much smaller than that of PC and MI.

From the confusion matrix, the values of FP, FN, sensitivity, and specificity are summarised in Table 9. Table 9 complements Table 5 by presenting the absolute numbers of errors (TN, FP, FN, TP) and specificity, which are not available in Table 5. In diagnostic evaluations, sensitivity and specificity are typically reported in pairs. Sensitivity (same as recall) is given in Table 5 as TP/(TP+FN), whereas specificity (TN/(TN+FP)) differs from precision (TP/(TP+FP)). Specificity is calculated against all true Normal cases, while precision is calculated against all AF-predicted cases. In the context of AF screening, FN (AF cases that were not detected) have greater clinical implications than FP (false alarms). With a reduction to 15 features, SHAP missed only 100 cases (FN), compared with PC, which missed 323, and MI, which missed 416. Thus, the sensitivity of SHAP-15 reached 95.62% and specificity 98.95%, far exceeding PC (85.85%) and MI (81.78%). With 20 features, all three methods had FN values: SHAP 92, PC 102, and MI 166; SHAP had the lowest FN and FP, with a sensitivity of 95.97% and a specificity of 98.82%.



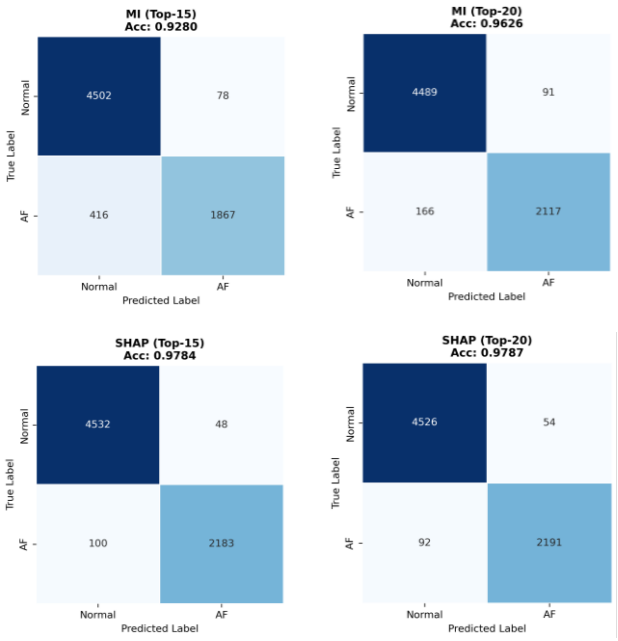


Figure 6. Confusion Matrix

TABLE 9. TABLE OF FALSE POSITIVE/FALSE NEGATIVE

Metric	15 Features			20 Features		
	MI	PC	SHAP	MI	PC	SHAP
TN	4502	4448	4532	4489	4494	4526
FP	78	132	48	91	86	54
FN	416	323	100	166	102	92
TP	1867	1960	2183	2117	2181	2191
Sensitivity (%)	81.78	85.85	95.62	92.73	95.53	95.97
Specificity (%)	98.30	97.12	98.95	98.01	98.12	98.82

The ROC curve can be seen in Figure 7. The SHAP curve is closest to the upper-left corner (ideal point: sensitivity = 1, 1-specificity = 0) and almost overlaps with the baseline curve, especially at 20 features, which visually confirms the results of DeLong's test that the AUC SHAP-20 does not differ significantly from the baseline.

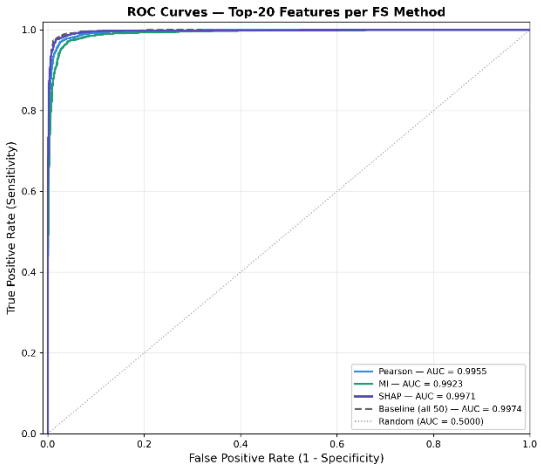


Figure 7. ROC Curves (Top 20 Features)

F. Features Analysis

Each feature selection method yields a different ranking, reflecting differences in evaluation methods. Table 10 lists the top five features for each method.

TABLE 10. TABLE OF TOP FIVE FEATURES OF EACH FEATURE SELECTION METHOD

Feature Selection	No	Features	Score
PC	1	R_Peak_ECG2	0.451
	2	R_Peak_ECG1	0.448
	3	RR_skew_ECG1	0.417
	4	RR_skew_ECG2	0.398
	5	rms_ECG2	0.382
MI	1	RR_std_ECG1	0.309
	2	RR_std_ECG2	0.270
	3	Delta_RR_std_ECG1	0.268
	4	Delta_RR_std_ECG2	0.221
	5	RR_mean_ECG1	0.211
SHAP	1	RR_std_ECG1	1.782
	2	RR_std_ECG2	0.642
	3	ECG2_Std_D2	0.592
	4	RR_skew_ECG2	0.411
	5	RR_mean_ECG1	0.360

Pearson Correlation ranks R_Peak_ECG2 (r=0.451) first because it has the highest linear correlation with the AF label. MI prioritises RR_std_ECG1 (MI=0.309), which captures R-R interval variability. SHAP also ranks RR_std_ECG1 first (score 1,782). These ranking differences reflect the characteristics of each feature selection method: Pearson captures linear relationships, MI measures information dependence, and SHAP evaluates feature contributions in the context of model interactions.

The visualisation of the SHAP ranking summarised in Table 10 comes from the SHAP Summary Plot (Figure 8). Based on the summary plot of the direction of influence of each feature value, the R-R interval variability feature, especially RR_std_ECG1 and RR_std_ECG2, ranks top. This is in accordance with the physiological characteristics of AF in the form of rhythmic irregularities. High R-R variability values tend to push predictions toward AF classes, while low values tend to push predictions toward Normal.

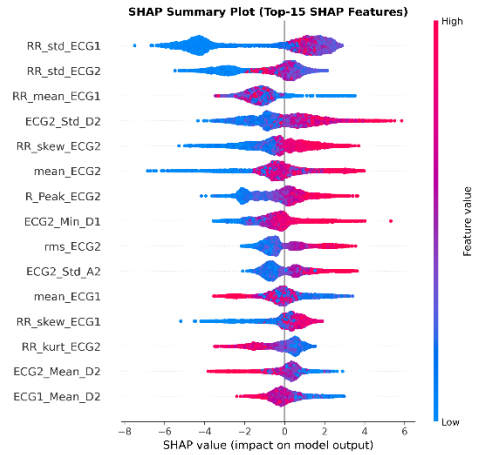


Figure 8. SHAP Summary Plot (Top 15)

For the 15 features selected by each feature selection method, the features common to all methods and those specific to each method are listed in Table 11. Four features are selected by the three (general) feature selection methods: R_Peak_ECG2, RR_skew_ECG1, ECG2_Std_A2, and RR_mean_ECG1. These four common features are drawn from each category: R_Peak_ECG2 from Morphological feature extraction, RR_skew_ECG1 and RR_mean_ECG1 from Statistical Characteristics feature extraction, and ECG2_Std_A2 from Wavelet Decomposition feature extraction. RR_skew_ECG1 represents the asymmetry of the distribution of the R-R interval as a marker of AF irregularity; RR_mean_ECG1 measures the average of the R-R interval; R_Peak_ECG2 captures the number of R peaks in ECG2; and ECG2_Std_A2 measures the variability of the SWT level 2 wavelet component in ECG2.

TABLE 11. TABLE OF FIFTEEN SELECTED FEATURES OF EACH FEATURE SELECTION METHOD

Features	Feature Selection			Descriptions
	PC	*MI*	*SHAP*	
R_Peak_ECG2	1	1	1	3 (General)
RR_skew_ECG1	1	1	1	3 (General)
ECG2_Std_A2	1	1	1	3 (General)
RR_mean_ECG1	1	1	1	3 (General)
ECG2_Std_D2	1		1	
RR_skew_ECG2	1		1	
rms_ECG2	1		1	
mean_ECG2		1	1	
RR_std_ECG1		1	1	
RR_std_ECG2		1	1	
R_Peak_ECG1	1	1		
RR_mean_ECG2	1	1		
ECG1_Mean_D2	1		1	
ECG2_Mean_D2	1		1	
ECG1_Std_A1	1			1 (Special PC)
ECG2_Min_A1	1			1 (Special PC)
ECG2_Std_A1	1			1 (Special PC)
rms_ECG1	1			1 (Special PC)
ECG2_Mean_A1		1		1 (Special MI)
ECG2_Mean_A2		1		1 (Special MI)
Delta_RR_mean_ECG1		1		1 (Special MI)
Delta_RR_mean_ECG2		1		1 (Special MI)
Delta_RR_std_ECG1		1		1 (Special MI)
Delta_RR_std_ECG2		1		1 (Special MI)
ECG2_Min_D1			1	1 (Special SHAP)
mean_ECG1			1	1 (Special SHAP)
RR_kurt_ECG2			1	1 (Special SHAP)
Total	**15**	**15**	**15**	

There are 4 special features in the PC feature selection, namely: ECG1_Std_A1, ECG2_Min_A1, ECG2_Std_A1, and rms_ECG1. In the selection of MI features, there are 6 special features, namely: ECG2_Mean_A1, ECG2_Mean_A2, Delta_RR_mean_ECG1, Delta_RR_mean_ECG2, Delta_RR_std_ECG1, and Delta_RR_std_ECG1 which capture the mean and variability of changes between consecutive R-R intervals in ECG1 and ECG2. Meanwhile, the special features of SHAP selection are ECG2_Min_D1, mean_ECG1, and RR_kurt_ECG2, which demonstrate SHAP's ability to capture information that other methods cannot.

IV. CONCLUSION

Based on the results of a comparative experiment with three feature selection methods (Pearson Correlation, Mutual Information, and SHAP) for classifying Atrial Fibrillation using XGBoost, the following conclusions were obtained. SHAP's feature selection outperformed PC and MI across all sizes (10, 15, and 20), with the largest difference at 10 features (SHAP 96.82% vs MI 79.40%; difference 17.42%). This advantage was tested statistically (McNemar and DeLong), showing that model-based post-hoc methods such as SHAP, calculated after XGBoost was fully trained and utilising the model's internal structure and inter-feature

interactions via TreeExplainer (model-specific), were more effective than model-free filter methods that evaluated features independently. Four features were selected by the three methods: R_Peak_ECG2 (Morphological), RR_skew_ECG1 (Statistical), ECG2_Std_A2 (Wavelet Decomposition), and RR_mean_ECG1 (Statistical). Features that set SHAP apart: ECG2_Min_D1, mean_ECG1, and RR_kurt_ECG2. Among the reduced configurations, SHAPs with 20 features achieved the highest performance (97.87% accuracy; F1 score 96.78%; ROC-AUC 99.71%; 60% reduction), while the 15-feature SHAP offers the best performance–dimensionality trade-offs: 97.84% accuracy, 97.85% precision, 95.62% recall, 96.72% F1-score, and 0.9965 ROC-AUC, with a 70%-dimension reduction. The difference between the two is very small (0.03% in accuracy), and even the SHAP-15 precision is higher (97.85% vs 97.59%). SHAP-15 differed by only 0.48% from the baseline accuracy with 50 features (98.32%). For ROC-AUC, only SHAPs with 20 features were not statistically different from the baseline (DeLong p=0.4997); in SHAP-15, the difference in AUC from the baseline (0.0009) is statistically significant but practically negligible. Thus, SHAP-15 offers a parsimonious–best-performing trade-off (70% reduction with minimal loss), while SHAP-20 guarantees ROC-AUC equivalence. The limitations of this study are that it uses only one dataset (MIT-BIH AFDB) without external validation, relies on a single classification algorithm (XGBoost), and selects the best configuration based on test-set performance. In addition, the proposed R-peak detection method (Adaptive Thresholding) has not been quantitatively compared with standard algorithms such as Pan-Tompkins and Hamilton.

REFFERENSI

[1] S. C. W. Tan, M.-L. Tang, H. Chu, Y.-T. Zhao, and C. Weng, "Trends in Global Burden and Socioeconomic Profiles of Atrial Fibrillation and Atrial Flutter: Insights from the Global Burden of Disease Study 2021," *CJC Open*, vol. 7, no. 3, pp. 247–258, Mar. 2025, doi: 10.1016/j.cjco.2024.11.017.

[2] V. Fuster *et al.*, "ACC/AHA/ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation," *Journal of the American College of Cardiology*, vol. 48, no. 4, pp. e149–e246, Aug. 2006, doi: 10.1016/j.jacc.2006.07.018.

[3] K. Miyazawa and G. Yh. Lip, "Atrial fibrillation," *Medicine*, vol. 46, no. 10, pp. 627–631, Oct. 2018, doi: 10.1016/j.mpmed.2018.07.009.

[4] R. Sharmin, M. C. Brindise, J. J. Kolliyil, B. A. Meyers, J. Zhang, and P. P. Vlachos, "Novel interpretable Feature set extraction and classification for accurate atrial fibrillation detection from ECGs," *Computers in Biology and Medicine*, vol. 179, p. 108872, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108872.

[5] D. A. Cook, S.-Y. Oh, and M. V. Pusic, "Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis," *JAMA Intern Med*, vol. 180, no. 11, p. 1461, Nov. 2020, doi: 10.1001/jamainternmed.2020.3989.

[6] S. L. Joshi, R. A. Vatti, and R. V. Tornekar, "A Survey on ECG Signal Denoising Techniques," in *2013 International Conference on Communication Systems and Network Technologies*, Gwalior: IEEE, Apr. 2013, pp. 60–64. doi: 10.1109/CSNT.2013.22.

[7] C. Guan *et al.*, "Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study," *Crit Care*, vol. 28, no. 1, p. 349, Oct. 2024, doi: 10.1186/s13054-024-05138-0.

[8] B. B.-S. Chuang and A. C. Yang, "Optimization of Using Multiple Machine Learning Approaches in Atrial Fibrillation Detection Based on a Large-Scale Data Set of 12-Lead Electrocardiograms: Cross-Sectional Study," *JMIR Form Res*, vol. 8, p. e47803, Mar. 2024, doi: 10.2196/47803.

[9] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, p. 88, Feb. 2023, doi: 10.3390/a16020088.

[10] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection".

[11] K. Cao *et al.*, "Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model," *Sci Rep*, vol. 15, no. 1, p. 12406, Apr. 2025, doi: 10.1038/s41598-025-96520-7.

[12] J. Li *et al.*, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017, doi: 10.1145/3136625.

[13] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *J Big Data*, vol. 11, no. 1, p. 44, Mar. 2024, doi: 10.1186/s40537-024-00905-w.

[14] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, "A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107865, May 2024, doi: 10.1016/j.engappai.2024.107865.

[15] J. R. Vergara and P. A. Estévez, "A Review of Feature Selection Methods Based on Mutual Information," *Neural Comput & Applic*, vol. 24, no. 1, pp. 175–186, Jan. 2014, doi: 10.1007/s00521-013-1368-0.

[16] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.

[17] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," *Clinical Translational Sci*, vol. 17, no. 11, p. e70056, Nov. 2024, doi: 10.1111/cts.70056.

[18] O. Heriana and A. M. Al Misbah, "Comparison of Wavelet Family Performances in ECG Signal Denoising," *indones.j.electron.telecommun.*, vol. 17, no. 1, p. 1, Aug. 2017, doi: 10.14203/jet.v17.1-6.

[19] S. Mandala, A. R. Pratiwi Wibowo, Adiwijaya, Suyanto, M. S. M. Zahid, and A. Rizal, "The Effects of Daubechies Wavelet Basis Function (DWBF) and Decomposition Level on the Performance of Artificial Intelligence-Based Atrial Fibrillation (AF) Detection Based on Electrocardiogram (ECG) Signals," *Applied Sciences*, vol. 13, no. 5, p. 3036, Feb. 2023, doi: 10.3390/app13053036.

[20] A. Géron, *Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, 3rd ed. O'Reilly Media, 2022.

[21] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985, doi: 10.1109/TBME.1985.325532.

[22] P. Hamilton, "Open source ECG analysis," in *Computers in Cardiology*, Memphis, TN, USA: IEEE, 2002, pp. 101–104. doi: 10.1109/CIC.2002.1166717.

[23] M. Shen, L. Zhang, X. Luo, and J. Xu, "Atrial Fibrillation Detection Algorithm Based on Manual Extraction Features and Automatic Extraction Features," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 428, no. 1, p. 012050, Jan. 2020, doi: 10.1088/1755-1315/428/1/012050.

[24] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994, doi: 10.1109/72.298224.

[25] A. Kraskov, H. Stoegbauer, and P. Grassberger, "Estimating Mutual Information," *Phys. Rev. E*, vol. 69, no. 6, p. 066138, Jun. 2004, doi: 10.1103/PhysRevE.69.066138.

[26] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] Q. A. Hidayaturrohman and E. Hanada, "A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes," *Applied Sciences*, vol. 15, no. 6, p. 3393, Mar. 2025, doi: 10.3390/app15063393.

[29] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.

[30] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.