

# Automated Detection and Classification of Arrhythmias in ECG Signals

Gowri Shankar A B

Department of Electronics and  
Communication

Amrita Vishwa Vidyapeetham  
Coimbatore

[cb.en.u4cce23001@cb.students.amrita.edu](mailto:cb.en.u4cce23001@cb.students.amrita.edu)

Dinesh Karthik V

Department of Electronics and  
Communication

Amrita Vishwa Vidyapeetham  
Coimbatore

[cb.en.u4cce23008@cb.students.amrita.edu](mailto:cb.en.u4cce23008@cb.students.amrita.edu)

Duvvuru Akshaya Saketh Reddy

Department of Electronics and  
Communication

Amrita Vishwa Vidyapeetham  
Coimbatore

[cb.en.u4cce23011@cb.students.amrita.edu](mailto:cb.en.u4cce23011@cb.students.amrita.edu)

Mugilan S S

Department of Electronics and  
Communication

Amrita Vishwa Vidyapeetham  
Coimbatore

[cb.en.u4cce23026@cb.students.amrita.edu](mailto:cb.en.u4cce23026@cb.students.amrita.edu)

**Abstract** — This paper presents an automated system for detecting and classifying cardiac arrhythmias from ECG data. The system utilizes machine learning techniques to analyze ECG signals, extract relevant features, and classify arrhythmias. The pipeline includes data preprocessing, feature extraction (QRS duration, R-R intervals, heart rate variability), and classification using supervised (Logistic Regression, Decision Tree, K-NN, Random Forests, SVM) and unsupervised (K-Means) learning algorithms.

**Keywords** — *Cardiac arrhythmias, Electrocardiogram (ECG), Machine learning, Data preprocessing, Feature extraction, Supervised learning, Unsupervised learning, Arrhythmia classification, Model evaluation, Accuracy, Precision, Recall, F1-score.*

## I. INTRODUCTION

Electrocardiograms (ECGs) are vital for diagnosing heart abnormalities, including arrhythmias, which affect heartbeat rate or rhythm. Manual ECG interpretation is time-consuming and error prone, especially with large datasets. This paper introduces, automated system for detecting and classifying arrhythmias using machine learning.

ECGs are fundamental tools in cardiology, providing a filtered representation of the heart's electrical activity. A typical heartbeat, as seen in a lead II ECG under normal sinus rhythm, exhibits a characteristic pattern comprising the P-wave (atrial depolarization), the QRS complex (ventricular depolarization), and the T-wave (ventricular repolarization). These waves are separated by segments like the PR, ST, and TP, and the intervals between and within these waves, such as the PR, QT, and RR intervals, are crucial diagnostic markers.

Damage to the heart muscle or nerves can change the electrical activity of the heart and induce a corresponding change in the shape of the ECGs. Thus, ECG is a major clinical diagnostic tool for various heart abnormalities.

The study classified ECG rhythms into five categories:

- Atrial fibrillation (AFib)
- Normal sinus rhythm (NSR)
- Sinus bradycardia (SB)
- Sinus tachycardia
- Supraventricular tachycardia (SVT)

## II. METHODOLOGY

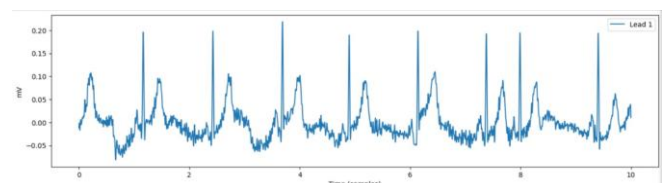
### A. Data Acquisition and Preprocessing

Our study utilized a novel dataset comprising 40,258 12-lead ECGs from 22,599 male and 17,659 female patients. These data were randomly selected from over 120,000 individuals who visited the Shaoxing People's Hospital and the Ningbo First Hospital of Zhejiang University between 2013 and 2018. The dataset includes 20% normal sinus rhythm (SR) and 80% abnormal readings. The majority of patients were between 51 and 80 years old.

The system ingests raw 12-lead ECG signals from **.mat** files, each representing a 10-second recording sampled at 500Hz.

Apply a four-stage filtering process:

- **Butterworth Low-Pass Filter:** Eliminate high-frequency noise above 50Hz, preserving crucial ECG components.
- **LOESS Smoothing:** Estimate and remove baseline wander using local regression.



- **Baseline Correction:** Centre the ECG signal around zero by subtracting the estimated baseline.
- **Non-Local Means Denoising:** Reduce random noise while maintaining signal features through patch-based similarity and exponential weighting.

This pipeline aims to enhance signal quality for subsequent analysis.

Why NLM is Used for ECG Denoising?

- **Preserves Morphological Features:** NLM excels at removing noise while retaining the crucial morphological components of the ECG signal, such as P-waves, QRS complexes, and T-waves, which are essential for accurate diagnosis.
- **Adaptive Smoothing:** Instead of applying uniform smoothing, NLM adaptively weights similar signal patches found across the entire ECG recording. This allows for effective noise reduction without overly smoothing or blurring sharp, clinically significant transitions within the signal.
- **No Need for Thresholding:** Unlike some other denoising techniques like wavelet denoising, NLM does not require the manual selection of thresholds, making it a more automated and potentially less subjective approach.

### B. Feature Extraction

A novel and interpretable feature extraction method was designed to capture relevant information from the preprocessed ECG signals. This method involved extracting both basic signal characteristics and more complex features related to the morphology and relationships between ECG waves (P-wave, QRS complex, T-wave) and segments.

A total of 11 distinct feature combinations were analyzed, ranging from a basic set of 11 characteristics to a comprehensive set of 39,830 features. Age and gender were included as features. For lead II ECG, features such as mean and variance of RR intervals, RR interval count, and statistical measures (mean, variance, prominence) of QRS complexes, non-QRS peaks, and valleys were extracted. Additionally, relationships between these peaks and valleys were quantified by calculating ratios of width difference, height difference, and prominence difference over time difference.

To handle the variable number of ratios per patient, an empirical frequency distribution table with 100 bins was constructed for these ratios and for the attributes of peaks and valleys. This approach allowed for the representation of complex morphological information in a fixed-length feature vector suitable for machine learning models.

### C. Classification and Evaluation

The extracted features were used to train and evaluate various machine learning classification models. An exhaustive grid search was performed to optimize the hyperparameters of classification algorithms, including Random Forest Classifier, K Nearest Neighbors (KNN), and Support Vector Machines (SVM).

The performance of the different classification schemes was evaluated on the preprocessed and feature-extracted datasets. Metrics such as accuracy, precision, recall, and F1-score were used to assess the robustness and reliability of the models in classifying the four rhythm groups (SB, AFIB, GSVT, SR). The impact of additional cardiac conditions on rhythm classification accuracy was also investigated by training and evaluating models on a subset of data without such conditions.

This systematic approach allowed for a comprehensive comparison of thousands of competing strategies to identify the optimal multi-stage arrhythmia classification routine.

### D. K-Means Clustering

The K-Means clustering implementation is a well-designed approach to uncover patterns in the ECG dataset. The process starts with data preparation, where the dataset is loaded using the `load_data` function. It selects only numeric columns to ensure compatibility with the clustering algorithm. The data is then standardized using `StandardScaler`, which adjusts each feature to have a mean of 0 and a standard deviation of 1. This standardization step is crucial because K-Means is sensitive to the scale of the features, and without it, the clustering results could be misleading. To determine the optimal number of clusters, the Elbow Method is applied. The algorithm evaluates different values for (the number of clusters) in the range of 2 to 10. For each value of `k`, the K-Means algorithm calculates inertia, which is the sum of squared distances between data points and their assigned cluster centres. The silhouette score, another metric for assessing clustering quality, is also computed for each value of `k`. This score measures how similar a data point is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. Both the Elbow Method and silhouette scores support the choice of `k=5` as the optimal number of clusters for this dataset.

### E. Kth Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is widely used for arrhythmia classification due to its simplicity, versatility, and effectiveness in handling such medical data. The process begins by loading the extracted features from CSV files, which contain ECG data and corresponding rhythm labels, respectively. The filenames in both datasets are cleaned to remove specific extensions for proper alignment. The datasets are then merged on matching filenames, with unnecessary columns being removed for simplification. Missing values in the merged dataset are handled by filling them with the median value of their respective columns, ensuring a consistent and clean dataset.

for model training. Features and target variables are separated, with the target variable () being encoded. This encodes the categorical rhythm labels into numeric values, allowing compatibility with machine learning algorithms. The dataset is then split into training and testing subsets, ensuring that 20% of the data is preserved for model evaluation. Stratification is applied during the split to maintain the class distribution. The features are standardized using, which transforms the data to have a mean of 0 and a standard deviation of 1. This preprocessing step ensures that all features have equal importance in the KNN algorithm, which is sensitive to feature magnitudes. As the dataset exhibits class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) method is applied to the training set. SMOTE generates synthetic examples for minority classes, balancing the distribution of classes and improving the model's ability to classify underrepresented categories. Hyperparameter tuning is conducted using to identify the optimal value for the number of neighbors () in the KNN algorithm. A range of values (1 to 20) is tested, and cross-validation is applied to evaluate the performance of each configuration. The best value for is selected based on the highest accuracy score during cross-validation.

The optimized KNN model is trained using the resampled training set. Predictions are made on the standardized testing set, and the model's performance is evaluated using metrics such as accuracy and the classification report. The classification report provides precision, recall, and F1-score for each class, along with overall accuracy, macro averages, and weighted averages.

#### *F. Decision Tree Classifier*

A Decision Tree is a hierarchical model that splits data based on feature values, creating branches that lead to classification outcomes. Decision Trees work well with both categorical and numerical data, making them suitable for arrhythmia classification where ECG features could include diverse types of data. The extracted ECG features are loaded using CSV file. Features and target variables are then separated. The target labels are encoded into numeric form, making them compatible with the machine learning algorithm. The data is split into training and testing subsets (80%-20%) with stratification applied to preserve the class distribution across both subsets. To ensure equal scaling of features, the StandardScaler() is employed. It adjusts each feature to have a mean of 0 and a standard deviation of 1. This standardization ensures that the Decision Tree's performance remains unaffected by differences in feature magnitudes. The Decision Tree classifier is then trained on the scaled training data. After training, the model predicts labels for the testing subset. Various evaluation metrics are calculated, including accuracy, a classification report, and a confusion matrix. These metrics offer insights into the model's performance.

#### *G. Random Forest Classifier*

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions for higher accuracy. It reduces overfitting by averaging

outputs, making it more robust than individual trees. It is widely used for classification tasks, especially in handling complex data structures with improved reliability.

Random Forest plays a crucial role in ECG rhythm classification by utilizing ensemble learning to improve accuracy and robustness. The extracted ECG features are loaded using CSV file. Categorical data within the features is encoded using one-hot encoding. This transformation ensures that non-numeric attributes are converted into a format that the Random Forest classifier can process. Additionally, feature selection is applied by selecting the most relevant features that contribute to rhythm classification. The dataset is split into training and testing subsets using stratified sampling to maintain class distribution. The Random Forest classifier is then trained on the selected features using RandomForestClassifier() with parameters such as n\_estimators=150, max\_depth=15, min\_samples\_split=5, and class\_weight='balanced'. These parameters optimize the model's ability to handle class imbalances and generalize well across different ECG rhythms.

#### *H. Regression*

Regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). Its primary goal is to predict or estimate the value of the dependent variable based on the given features. Regression techniques are widely used in fields such as finance, healthcare, and environmental studies to analyze trends and make predictions.

Logistic Regression models the probability of an arrhythmia occurrence using the sigmoid function. It outputs values between 0 and 1, where thresholds (e.g., 0.5) determine whether a rhythm is classified as normal or arrhythmic. It is trained with a balanced class weight to address class imbalances in arrhythmia data. The trained model evaluates predictions on the testing set, achieving metrics like accuracy, precision, recall, F1-score, and AUC-ROC. These metrics demonstrate the model's effectiveness in distinguishing between normal and arrhythmia classes. A confusion matrix is visualized to identify classification strengths and errors, while feature importance is analyzed to highlight the most influential ECG parameters for arrhythmia prediction.

#### *I. Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal hyperplane that separates classes in a feature space. It uses mathematical functions (kernels) to transform data, enabling it to classify non-linear patterns effectively. SVM handles high-dimensional data well and provides robust performance, especially with smaller datasets.

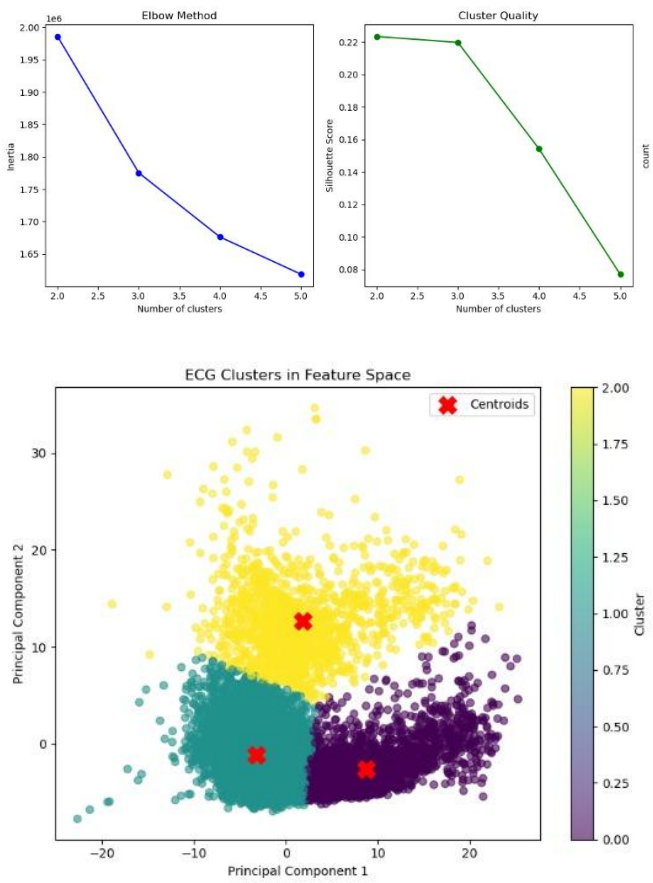
SVM is ideal for arrhythmia classification due to its ability to distinguish subtle differences in ECG features. Pre-processed ECG data, including clinically relevant features like heart rate variability (HRV) and intervals, is fed into the algorithm.

SVM uses advanced kernels, like the radial basis function (RBF), to model complex relationships between features and arrhythmia class. Balanced class weights address imbalances in arrhythmia data, ensuring effective classification of both common and rare arrhythmia types. GridSearchCV optimizes hyperparameters, enhancing the model's accuracy. Metrics like precision, recall, and F1-score evaluate the model, while confidence analysis identifies prediction reliability. SVM's high-dimensional feature handling and robust classification make it a powerful tool for diagnosing arrhythmias.

III. RESULT

A. K-Means Clustering

The analysis supports k=3 as the best choice for clustering.



The scatter plot validates the separation of clusters, useful for identifying ECG data patterns (e.g., normal vs. abnormal signals).

B. Kth Nearest Neighbors (KNN)

The model achieves an accuracy of 75%, indicating its ability to classify ECG rhythms reasonably well.

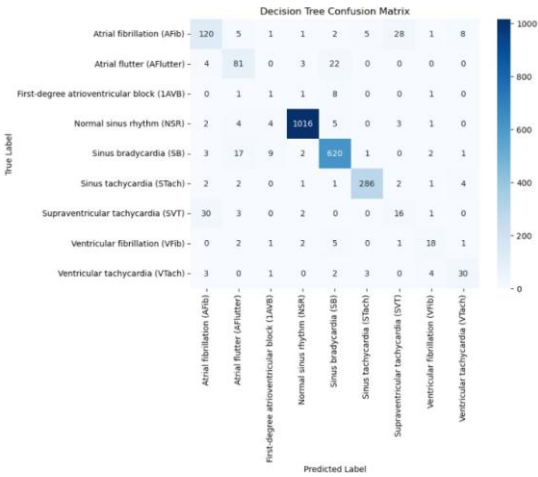
Best k: 1				
Accuracy: 0.75				
Classification Report:				
	precision	recall	f1-score	support
Atrial fibrillation (AFib)	0.66	0.64	0.65	171
Atrial flutter (AFlutter)	0.40	0.76	0.52	110
First-degree atrioventricular block (IABV)	0.29	0.58	0.39	12
Normal sinus rhythm (NSR)	0.95	0.73	0.82	1035
Sinus bradycardia (SB)	0.69	0.77	0.73	655
Sinus tachycardia (STach)	0.85	0.92	0.89	299
Supraventricular tachycardia (SVT)	0.17	0.21	0.19	52
Ventricular fibrillation (VFib)	0.37	0.63	0.47	30
Ventricular tachycardia (VTach)	0.81	0.79	0.80	43
accuracy			0.75	2407
macro avg	0.58	0.67	0.61	2407
weighted avg	0.79	0.75	0.76	2407

The optimal number of neighbors (k) chosen is 1, determined through hyperparameter tuning. Using too few neighbors makes the model memorize training data. More neighbors improve performance, but only up to a point.

C. Decision Tree Classifier

The model achieves an accuracy of 91%.

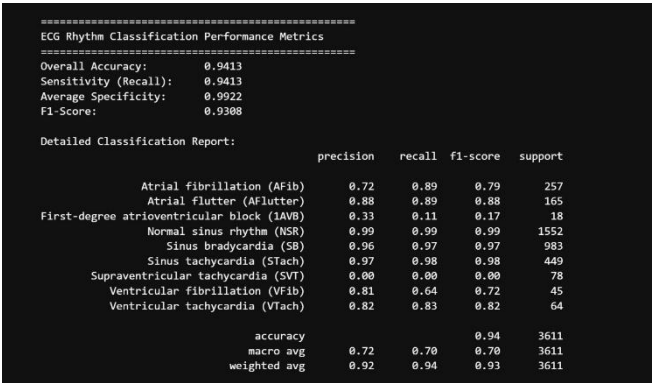
Accuracy: 0.91				
Classification Report:				
	precision	recall	f1-score	support
Atrial fibrillation (AFib)	0.73	0.70	0.72	171
Atrial flutter (AFlutter)	0.70	0.74	0.72	110
First-degree atrioventricular block (IABV)	0.06	0.08	0.07	12
Normal sinus rhythm (NSR)	0.99	0.98	0.98	1035
Sinus bradycardia (SB)	0.93	0.95	0.94	655
Sinus tachycardia (STach)	0.97	0.96	0.96	299
Supraventricular tachycardia (SVT)	0.32	0.31	0.31	52
Ventricular fibrillation (VFib)	0.62	0.60	0.61	30
Ventricular tachycardia (VTach)	0.68	0.70	0.69	43
accuracy			0.91	2407
macro avg	0.67	0.67	0.67	2407
weighted avg	0.91	0.91	0.91	2407



The confusion matrix shows accurate classification for dominant classes like NSR, while minority classes such as VFIB and SVT have more misclassifications, which is typical with imbalanced datasets. The Decision Tree classifier excels in classifying prevalent rhythms such as NSR and SB with high precision and recall.

D. Random Forest Classifier

The Random Forest Classifier achieved an excellent overall accuracy of 94.13%, showcasing its strength in classifying ECG rhythms with high reliability. It also demonstrated a sensitivity (recall) of 94.13%, indicating its ability to correctly identify positive cases across rhythm classes. Additionally, the model attained an average specificity of 99.22%, reflecting its effectiveness in minimizing false positives for non-target classes. The weighted F1-score of 93.08% underscores a strong balance between precision and recall, making it a robust choice for arrhythmia classification.

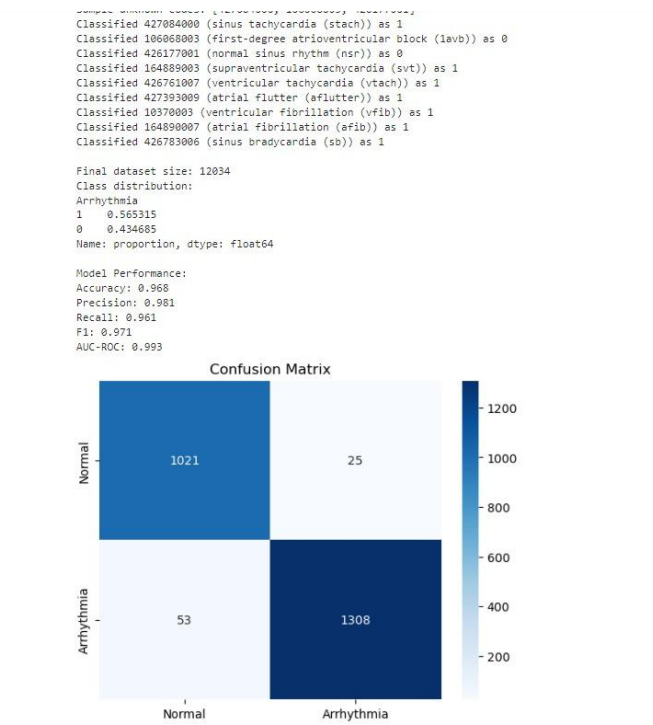


The Confusion matrix provides valuable insights into the model's performance. For common rhythms like Normal Sinus Rhythm (NSR) and Sinus Bradycardia (SB), the diagonal cells show strong classification accuracy with minimal errors. However, minority rhythms like Supraventricular Tachycardia (SVT) and Ventricular Fibrillation (VFib) exhibit more misclassifications, suggesting challenges in handling less frequent arrhythmias.

E. Regression

The Logistic Regression model achieved 96.8% accuracy, indicating strong reliability in classifying ECG rhythms. The precision of 98.1% confirms that the model effectively minimizes false positives, while the recall of 96.1% shows its ability to correctly identify arrhythmic cases. The F1-score of

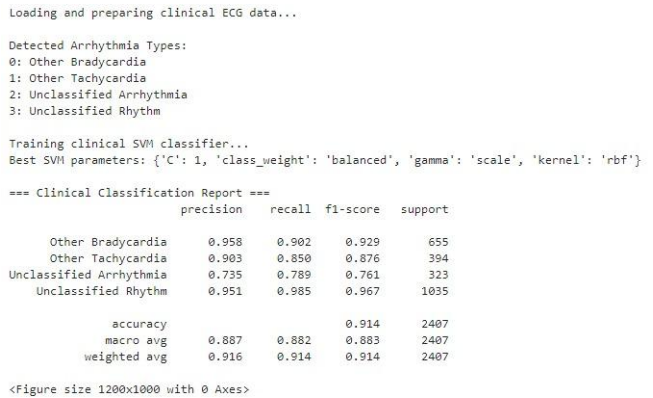
97.1% demonstrates an excellent balance between precision and recall. Furthermore, the AUC-ROC score of 99.3% signifies outstanding model performance in distinguishing normal and arrhythmic conditions.



The confusion matrix highlights the classification results, where Normal Sinus Rhythm (NSR) is correctly classified in 1,021 cases, with only 25 misclassified as arrhythmia. Meanwhile, arrhythmia cases show 1,308 correct classifications, with 53 misclassified as normal rhythms. This distribution suggests that the model effectively recognizes most ECG rhythms, though a minor margin of error exists.

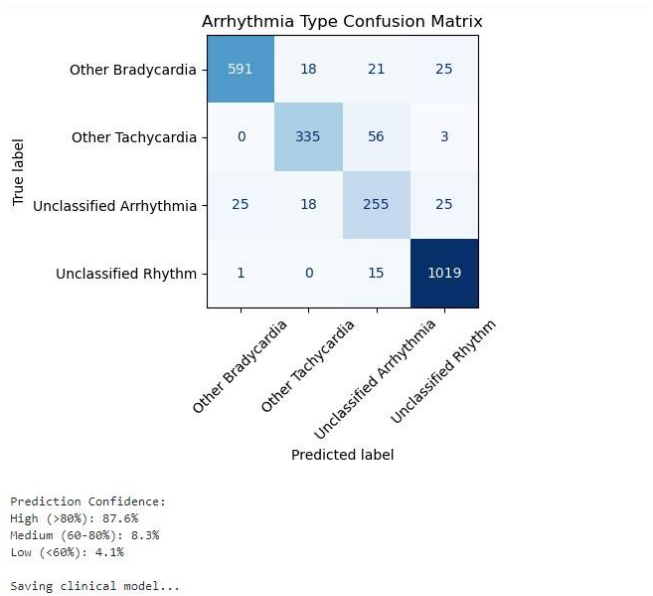
F. Support Vector Machine (SVM)

The Support Vector Machine (SVM) achieved an overall accuracy of 91.4%, demonstrating strong performance in classifying ECG rhythms. The macro-average precision of 88.7% and recall of 88.2% suggest that the model effectively differentiates between rhythm types while maintaining a balanced classification approach. The weighted-average precision of 91.6% ensures minimal false positives.





Breaking down the classification results, Other Bradycardia was classified with 95.8% precision and 90.2% recall, indicating strong accuracy in detecting this rhythm type. Similarly, Other Tachycardia achieved 90.3% precision and 85.0% recall, confirming the model's ability to identify tachycardic conditions effectively. The Unclassified Arrhythmia category exhibited lower precision at 73.5%, but recall remained high at 78.9%, showing moderate difficulty in handling ambiguous rhythms. The Unclassified Rhythm achieved outstanding metrics, with 95.1% precision and 98.5% recall, demonstrating reliable identification of known ECG patterns.



591 instances of Other Bradycardia were correctly classified, while 18 were misclassified as Other Tachycardia, 21 as Unclassified Arrhythmia, and 25 as Unclassified Rhythm. This suggests strong performance for common rhythm types, but some confusion remains between similar categories. The overall prediction confidence percentages indicate that 87.6% of classifications had high confidence (>80%), while 8.3% fell within medium confidence (60-80%), and 4.1% had low confidence (<60%), highlighting the model's reliability.

IV. CONCLUSION

This study presented an automated system for detecting and classifying cardiac arrhythmias using machine learning techniques applied to 12-lead ECG signals. The proposed pipeline—comprising preprocessing, feature extraction, and classification—demonstrated robust performance across multiple models, with Logistic Regression achieving the highest accuracy (96.8%), followed by Random Forests (94.13%) and SVM (91.4%). Key findings include:

Effective Feature Extraction: The novel combination of temporal (RR, PR, QT intervals) and morphological (QRS, P/T waves) features, alongside HRV metrics, enabled discriminative representation of arrhythmias.

Model Performance:

Supervised models (e.g., Random Forest) excelled in classifying common rhythms (NSR, SB) but faced challenges with minority classes (SVT, VFib).

Unsupervised clustering (K-Means, k=3) validated the separability of ECG patterns, aiding exploratory analysis.

REFERENCES

[1] J. Zheng, H. Chu, D. Struppa, et al., "Optimal Multi-Stage Arrhythmia Classification Approach," Scientific Reports, vol. 10, no. 2898, pp. 19 February 2020.

[2] H. K. Kim and M. H. Sunwoo, "An Automated Cardiac Arrhythmia Classification Network for 45 Arrhythmia Classes Using 12-Lead Electrocardiogram," in IEEE Access, vol. 12, pp. 44527-44538, 2024.

[3] N. S, J. M, I. T, S. R and K. G, "Pre-Cardiac arrhythmia detection using Machine Learning," 2023 International Conference on Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM), Coimbatore, India, 2023