Atrial Fibrillation Detection Using Multi-Model Approaches

1st Chandini Karrothu
Dept. of Computer Science
Kent State University
Kent, USA
ckarroth@kent.edu

2nd Likhitha Marrapu
Dept. of Computer Science
Kent State University
Kent, USA
lmarrapu@kent.edu

3rd Shivani Battu
Dept. of Computer Science
Kent State University
Kent, USA
sbattu1@kent.edu

***Abstract*—Atrial Fibrillation (AF), the most common cardiac arrhythmia, poses significant risks for stroke and heart failure, underscoring the critical need for accurate and early detection. This study focuses on classifying AF using the Physionet 2017 database - AF classification from a short single lead ECG recording, a benchmark dataset containing four distinct ECG signal classes, leveraging advanced machine learning techniques to address the inherent challenges of signal variability, noise, and class imbalance. Models such as Dual Support Vector Machines (SVM) with linear and RBF kernels, LightGBM, Convolutional Neural Networks (CNN), and a Hybrid model combining SVM and LightGBM were implemented and compared.**

The Hybrid model achieved highest performance, with an accuracy of 76.15%, precision of 74.94%, recall of 76.15%, and F1 score of 74.7%, demonstrating the effectiveness of combining complementary model strengths. While LightGBM showed robust individual performance, CNN and Dual SVM struggled with precision-recall trade-offs, particularly for minority classes. Preprocessing steps, such as sliding window segmentation and signal inversion correction, mitigated dataset challenges and improved analysis reliability. Confusion matrix insights highlighted the Hybrid model's balanced classification across all classes, outperforming individual models in handling imbalanced data.

This study determines the value of hybrid models in enhancing AF detection accuracy and robustness, paving way for more reliable diagnostic tools. Future work can explore advanced techniques like cost-sensitive learning, data augmentation, and domain-specific feature engineering to further improve performance in complex, imbalanced classification tasks.

***Index Terms*—Atrial Fibrillation, Support Vector Machine, LightGBM, Signal Noise, Signal Entrophy**

I. Introduction

Atrial Fibrillation (AF) is the most prevalent cardiac arrhythmia, affecting millions of individuals worldwide and posing significant risks for stroke, heart failure, and other cardiovascular complications [4]. Early and accurate detection of AF is critical for effective treatment and management, which can significantly reduce mortality and morbidity rates [7]. Electrocardiogram (ECG) signals, being the primary diagnostic tool for AF, provide a rich source of information but present challenges in automated analysis due to signal variability, noise, and overlapping characteristics with other arrhythmias [5]. The complexity of ECG signal interpretation requires advanced computational approaches to improve diagnostic accuracy and reliability. This study addresses the classification of Atrial Fibrillation using the CINC2017 database [2], a widely used benchmark dataset containing four distinct classes of ECG signals. The choice of AF as the focus of this project stems from its clinical importance, as well as the challenges it presents for automated classification. AF detection is particularly challenging due to its episodic nature and subtle variations in ECG waveforms, making it an ideal candidate for exploring advanced machine learning techniques. To tackle this problem, we employed a range of machine learning models, including Dual Support Vector Machines (SVM) with linear and radial basis function (RBF) kernels, LightGBM, Convolutional Neural Networks (CNNs), and a Hybrid model combining SVM and LightGBM. The Hybrid model demonstrated superior performance across all metrics, highlighting the effectiveness of leveraging complementary model strengths. While CNNs showed promise, their relatively lower performance emphasized the need for approaches better suited to time-series data like ECG signals. Our approach also addressed key challenges inherent in the CINC2017 dataset [2], such as signal length variations, the limitations of fixed bandpass filters, feature selection biases, and class imbalances. For instance, a sliding window technique was adopted to manage varying signal durations, and careful preprocessing, including signal inversion correction, was performed to ensure the accuracy of the analysis. The results of our comparative analysis demonstrate the potential of hybrid models in improving AF detection accuracy, achieving a peak performance with an accuracy of 76.15%, precision of 74.94%, recall of 76.15%, and F1 score of 74.7%. This study underscores the value of combining traditional machine learning algorithms with modern techniques to address the complexities of ECG signal classification, paving the way for more robust diagnostic tools in clinical practice.

II. Related Work

Atrial Fibrillation (AF) detection using machine learning and deep learning methods has seen substantial progress in recent years, with advancements focusing on improving classification accuracy and addressing challenges like noise, data imbalance, and feature extraction. The below significant studies provided foundational insights and methodologies that guided the present work:

Geweid and Chen (2022) proposed a Hybrid Approach of Dual Support Vector Machine (HA-DSVM) for the classification of AF from short single-lead ECG recordings. Their methodology used a combination of standard SVM and a dual-SVM to enhance the classification accuracy. The process included signal decomposition using Discrete Wavelet Transform (DWT), which effectively addressed noise and baseline drift issues, and avoided the need for manual feature extraction. Their results, validated on the PhysioNet 2017 Challenge dataset, demonstrated high reliability, with an F1 score of 0.95 and an accuracy of 99.27% on the validation set, showcasing the robustness of their hybrid approach for multi-class classification [3].

Wang et al. (2023) introduced an end-to-end Dual-Path Recurrent Neural Network (DPRNN) for AF detection, leveraging the PhysioNet 2017 Challenge dataset. Their model segmented ECG signals into shorter overlapping windows for intra- and inter-segmental modeling, addressing the sequential nature of ECG data. The addition of mix-up data augmentation effectively mitigated issues related to limited data availability and model overfitting. This approach outperformed conventional feature-based machine learning methods and state-of-the-art deep learning baselines, demonstrating the strength of DPRNN for ECG-based arrhythmia detection [6].

These studies highlighted the effectiveness of hybrid and end-to-end modeling techniques in addressing the unique challenges posed by single-lead ECG data, such as variability in signal quality and class imbalance. Building on these approaches, the present work combines machine learning and hybrid techniques to enhance the detection of AF, with preprocessing steps like noise reduction, signal inversion correction, and feature extraction tailored to the dataset's characteristics.

Chuang et al [1] explores the use of machine learning to improve atrial fibrillation detection. By analyzing power spectral features from ECG recordings, it identifies the LightGBM model as the most effective, achieving an average F1-score of 0.988. The study underscores the clinical benefits of machine learning in real-time detection, particularly in intensive care and remote monitoring setups, demonstrating its potential to enhance patient care outcomes.

III. Dataset Overview

The PhysioNet/Computing in Cardiology Challenge 2017 dataset [2] is a comprehensive collection of single short-lead ECG recordings, ranging from 30 to 60 seconds in duration recorded at a 300 Hz sampling rate and band-pass filtered for noise reduction. These recordings were labeled into four distinct classes: Normal Sinus Rhythm (5,154 recordings), Atrial Fibrillation (771 recordings), Other Rhythm (2,557 recordings), and Noisy Recordings (46 recordings). The dataset was developed to advance the capabilities of automatic cardiac rhythm classification systems, addressing the significant challenge of detecting AF, which is often episodic and can be confused with other arrhythmic patterns.

The recordings were sourced from AliveCor devices, capturing real-world conditions including noise and variability. The dataset provides a benchmark for researchers to test algorithms capable of distinguishing AF from other cardiac rhythms and noise. It combines atrial activity analysis and ventricular response methods, leveraging features such as P-wave absence and RR interval irregularity. This multi-faceted approach enables the development of algorithms that can handle the complexities of real-world cardiac data.

By supporting algorithmic development for accurate and robust classification, this dataset addresses the global health challenge posed by AF, which is associated with increased risks of stroke and heart failure. The dataset serves as a vital resource for enhancing diagnostic tools, particularly for wearable and real-time monitoring systems. It not only emphasizes classification accuracy but also prioritizes resilience to noise and irregularities, fostering advancements in clinical and remote healthcare solutions.

IV. Methodology

A. Preprocessing Flow

The preprocessing of ECG signals was performed in a series of steps to ensure the data's suitability for analysis. These steps are outlined as follows:

1) Loading the ECG Signal: Raw ECG signals were loaded from `.mat` files using the `scipy.io.loadmat()` function, enabling structured data handling.

2) Filtering: A bandpass filter (0.5–50 Hz) was applied to the raw signals (Figure 1) to remove noise and artifacts while retaining the relevant frequency components essential for analysis. This resulted in a filtered signal as in Figure 2

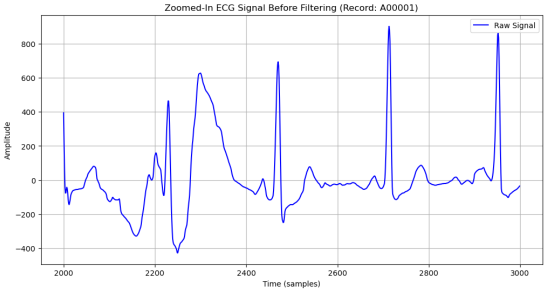

Fig. 1. Original AF ECG Signal

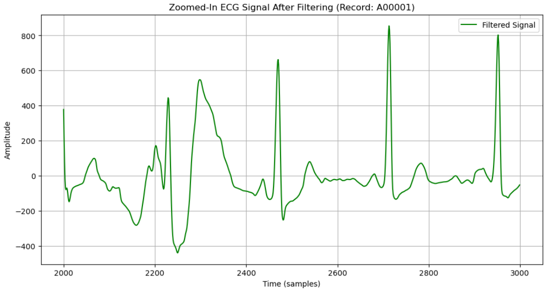

Fig. 2. Filtered ECG Signal

3) Normalization: The ECG signals were standardized using Z-score normalization to achieve zero mean and unit variance. This process ensured comparability across signals and improved the stability of downstream models.

4) Signal Correction: Inverted signals were corrected based on their polarity and peak characteristics. This step ensured accurate representation of the signal for further feature extraction.

5) Segmentation: The signals were divided into overlapping segments using a sliding window of 10 seconds with a 5-second overlap. This segmentation allowed local features to be extracted from different parts of the signal, enhancing the robustness of the analysis.(Figure 3)

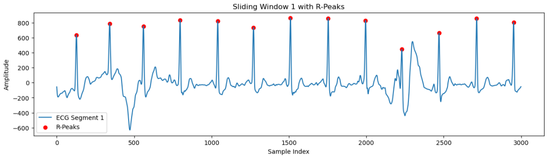

Fig. 3. Sliding Window with R peaks

6) Feature Extraction: A diverse set of features was extracted from each segmented signal such as Time-Domain Features, Statistical Features, Frequency-Domain Features, Poincare Features.

The following features were extracted from the ECG signals to assist in rhythm classification:

- **mean_rr**: The average duration between consecutive R-wave peaks in the ECG signal, which is crucial for assessing heart rate and rhythm.
- **sdnn**: The standard deviation of RR intervals, a time-domain measure of heart rate variability (HRV). Low SDNN values are often associated with increased risk for arrhythmias, including atrial fibrillation.
- **rmssd**: The root mean square of successive RR interval differences, another time-domain measure of HRV, indicating beat-to-beat variability. Low RMSSD values suggest disturbed autonomic function, which could indicate arrhythmia.
- **mean**: The average value of the ECG signal over its entire duration. This feature helps in detecting baseline shifts that may indicate electrical activity problems.
- **std_dev**: The standard deviation of the ECG signal, which measures the spread or variability in amplitude and can be used to detect irregularities like atrial fibrillation.
- **skewness**: A measure of the asymmetry of the ECG signal's distribution. Positive skewness suggests the signal has a tail on the right side, while negative skewness suggests a left-side tail. This feature is useful for identifying abnormal heart rhythms.
- **kurtosis**: A statistical measure of the ”tailedness” of the ECG signal's distribution. High kurtosis values may indicate abnormal events in the ECG, such as arrhythmic peaks.
- **dominant_freq**: The frequency at which the largest spectral power occurs in the ECG signal. It is particularly useful for identifying characteristic frequencies linked to irregular heart rhythms like atrial fibrillation.
- **total_power**: The total energy contained within the ECG signal over the entire frequency spectrum. Higher values indicate increased heart rate variability, often associated with conditions like atrial fibrillation.
- **poincare_sd1**: The standard deviation of the short-term axis of the Poincaré plot, which reflects high-frequency HRV. This feature provides insight into the heart's beat-to-beat variability.
- **poincare_sd2**: The standard deviation of the long-term axis of the Poincaré plot, reflecting low-frequency HRV. It is used to assess overall heart rate variability and can indicate autonomic dysfunction, common in atrial fibrillation.
- **poincare_ratio**: The ratio of SD1 to SD2 on the Poincaré plot. This ratio helps differentiate between types of arrhythmias and is useful for detecting atrial fibrillation.
- **max_amplitude**: The maximum amplitude of the ECG signal, which indicates the intensity of electrical activity in the heart. Significant changes in amplitude can be linked to arrhythmias.
- **mean_amplitude**: The average amplitude of the ECG signal over time, which provides an overview of the heart's electrical signal strength. Variations in amplitude can indicate arrhythmias like atrial fibrillation.
- **amplitude_sd**: The standard deviation of the ECG signal's amplitude. This feature measures variability in the signal's intensity and is useful for detecting irregularities in heart rhythms.
- **signal_entropy**: A measure of the complexity or randomness of the ECG signal. A higher entropy indicates more irregularity, which is characteristic of conditions like atrial fibrillation.
- **signal_energy**: The total energy contained within the ECG signal. Changes in energy levels can indicate abnormalities in heart function and rhythm, such as those caused by atrial fibrillation.
- **label**: The ground truth or classification label, which denotes whether the ECG signal corresponds to a normal rhythm or a condition like atrial fibrillation.

7) Labeling and Storage: Each segmented signal was assigned a corresponding label—Normal sinus rhythm, Atrial Fibrillation, Other Rhythm, or Too Noisy—and the features were stored in a structured DataFrame for subsequent machine learning tasks.

8) Missing Values: Missing values in the dataset, particularly in the poincare_sd1, poincare_sd2, and poincare_ratio columns, were imputed using the median of each column's available data. This method was chosen as the median is less sensitive to outliers, ensuring more robust imputation without distorting the data distribution.

9) Scaling the Features: The use of Min-Max normalization scales the numeric features within a defined range, typically between 0 and 1. This transformation ensures that all features contribute equally to the model by eliminating any bias introduced by differing scales or magnitudes. By normalizing the data, extreme values are compressed, preventing features with larger numerical ranges from dominating the analysis.

10) Label Encoding: The label encoding process has converted the categorical class labels into numeric values to facilitate machine learning model training. In this mapping, the label ’A’ (Atrial Fibrillation) is encoded as 0, ’N’ (Normal rhythm) as 1, ’O’ (Other rhythm) as 2, and ’˜’ (Noisy signal) as 3. This transformation ensures that the target variable is in a numerical format, enabling algorithms to process it effectively without implying any ordinal relationship between the categories.

This preprocessing pipeline ensured high-quality, noise-free data for robust classification and feature analysis.

B. Exploratory Data Analysis

1) Class Imbalance: The label distribution (Figure 4) shows an imbalance in the dataset, with the majority of instances belonging to the ’N’ (Normal) class, which has 27,039 samples. This is followed by the ’O’ (Other) class with 13,939 samples. The ’A’ (Atrial Fibrillation) class has 4,075 samples, and the ’ ’ (Noisy) class has the fewest samples at 1,012. This uneven distribution suggests that the model may be more exposed to normal and other types of signals, while the Atrial Fibrillation and noisy classes are underrepresented, potentially leading to challenges in detecting rarer events like atrial fibrillation.

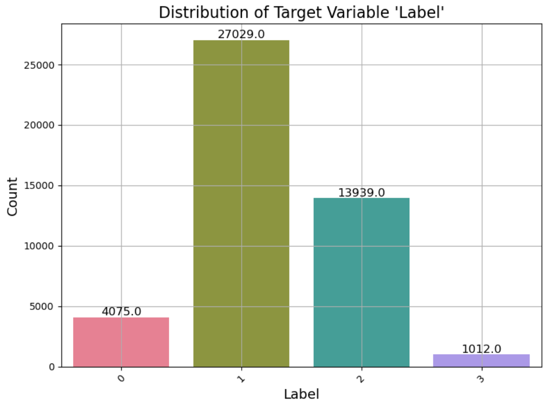

Fig. 4. Label Distribution in the Target Attribute

C. Correlation Analysis

The correlation matrix (Figure 5) presented below reveals the relationships between various ECG-derived features. Strong positive correlations are observed between related features, such as ”Mean RR Interval” and ”HRV,” highlighting the inverse relationship between heart rate and the mean RR interval. ”SDNN,” ”RMSSD,” and ”Poincare SD1” also show strong positive correlations, indicating that these measures of heart rate variability are interconnected. The ”QRS Duration” feature is highly negatively correlated with ”HRV,” which is expected, as longer QRS durations are often associated with lower heart rate variability. Additionally, features like ”Total Power” and ”LF Power” show strong correlations with each other, suggesting they provide similar information about the signal’s frequency domain characteristics. This matrix provides valuable insights for feature selection, as highly correlated features may be redundant for model training.

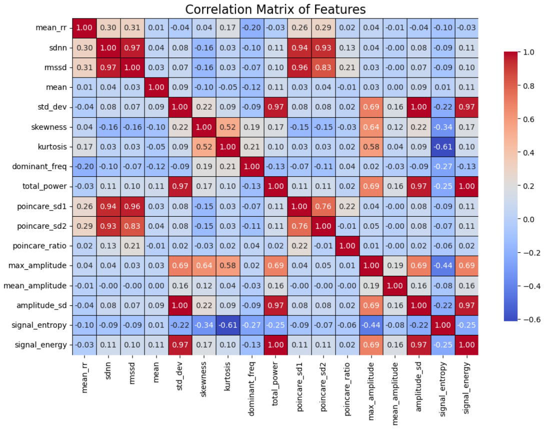

Fig. 5. Correlation Matrix of the attributes

D. Univariate Analysis

The histograms shown below (Figure 6) represent the distribution of various features extracted from the ECG signals. Most of the features display a highly skewed distribution, with many of them having peaks concentrated at lower values, such as ”mean_rr,” ”sdnn,” and ”rmssd.” This indicates that the majority of the data points fall within a small range of values, while only a few observations deviate significantly from this range. Features like ”poincare_sd1,” ”signal_entropy,” and ”signal_energy” exhibit similar patterns, suggesting that the data for these features are concentrated around certain values with fewer extreme outliers. These insights into the distributions can inform the preprocessing strategy, especially for normalization or transformation methods to enhance model performance and prevent biases due to skewed data.

1) Bivariate Analysis: This visualization provides a clear view of the distribution of various features across different labels through density plots. Each plot compares the feature distribution of the classes (denoted by labels) to showcase how distinct or overlapping the distributions are. For most features, such as mean_rr, mean_amplitude, and signal_entropy, the class distributions seem to differ significantly, indicating that these features may be important for distinguishing between classes. However, some features like poincare_sd1 and dominant_freq exhibit overlapping distributions, suggesting that they might be less effective in separating the classes.

2) Stratified Random Sampling: The dataset has been split using an 80:20 ratio, with 36,844 samples allocated for training and 9,211 samples for testing. Stratified random sampling was employed to preserve the class distribution in both the training

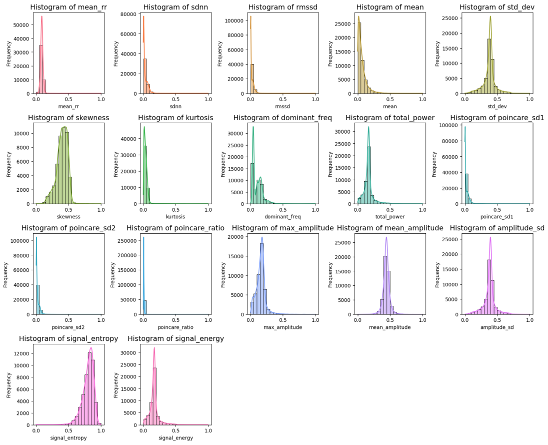

Fig. 6. Distribution of the Predictor variables

and testing sets, addressing potential class imbalances depicted in Figure 7. This ensures that each class is proportionally represented in both sets, facilitating more reliable model training and evaluation by preventing skewed predictions towards the majority class. The consistent distribution across classes in both splits will help improve the model's generalization and accuracy, particularly for underrepresented classes.

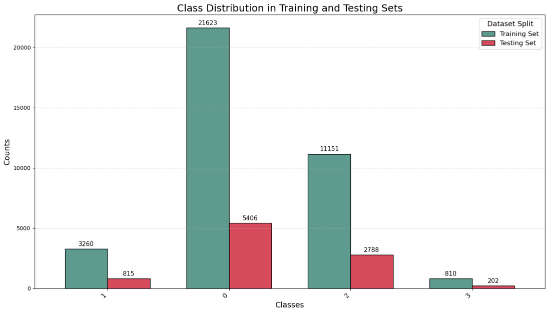

Fig. 7. Class Distribution in the Training and Testing sets

E. Models

1) Dual Support Vector Machines: Dual Support Vector Machine (SVM) is implemented with hyperparameter tuning for classification using linear and radial basis function (RBF) kernels. SVM is chosen due to its ability to handle high-dimensional data and its robust performance in classification tasks.

Model Initialization: Two SVM models are defined with linear and RBF kernels. These kernels allow the algorithm to separate classes in a linear and non-linear manner, respectively. Both models are initialized with `random_state=42` for reproducibility.

Feature Scaling: As SVMs are sensitive to feature scaling, the dataset is standardized using `StandardScaler`. Scaling ensures that features contribute equally to the model's decision boundary by transforming them to a uniform scale with a mean of zero and unit variance.

Hyperparameter Tuning: To optimize the performance of the SVM models, hyperparameter tuning is conducted using `GridSearchCV`. The following parameters are tuned:

- **C:** Regularization parameter to control the trade-off between achieving a low error on training data and minimizing model complexity.
- **gamma:** Kernel coefficient influencing the decision boundary's shape in RBF kernels.

A 3-fold cross-validation is used for evaluation, ensuring robust and generalized results.

Model Training: After determining the optimal hyperparameters, the SVM models are retrained on the scaled training data using the best parameter configurations obtained from `GridSearchCV`.

Prediction and Ensemble: Predictions are made on the test dataset for both linear and RBF models. A simple ensemble method—majority voting—is applied to combine the predictions, improving overall classification performance by leveraging the strengths of both models.

Evaluation Metrics: The combined predictions are evaluated using standard classification metrics:

- **Accuracy:** Measures the proportion of correctly classified samples.
- **Precision:** Assesses the model's ability to avoid false positives.
- **Recall:** Evaluates the model's ability to identify all positive samples.
- **F1-Score:** Balances precision and recall, especially useful for imbalanced datasets.

A classification report provides detailed performance metrics.

Confusion Matrix: The confusion matrix is computed to visualize the model's performance across classes. It highlights true positives, false positives, false negatives, and true negatives, offering deeper insights into classification errors.

ROC Curve: To visualize the model's performance, ROC curves were plotted for each class. The curves (Figure 8) demonstrate the trade-off between true positive and false positive rates. The area under each curve (AUC) was calculated to quantify the model's discrimination ability. Figure illustrates the ROC curves for all classes.

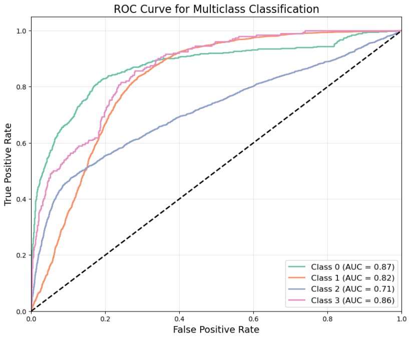

Fig. 8. ROC Curve for Dual SVM

This systematic approach ensures the implementation of optimized SVM models and provides comprehensive performance evaluation, making the methodology robust and reproducible for similar classification tasks.

2) LightGBM: LightGBM, a gradient-boosting framework, was implemented for multiclass classification. Its efficiency in handling large datasets and ability to model complex relationships made it a suitable choice for this task. The methodology is outlined below:

Dataset Preparation: The training and testing datasets were converted into LightGBM-compatible formats using the `lgb.Dataset` function, ensuring compatibility with the framework.

Hyperparameter Tuning: A grid search was performed to optimize the following parameters:

- **num_leaves:** Controls tree complexity and enhances learning capacity.
- **learning_rate:** Adjusts the contribution of each tree to the overall model.
- **max_depth:** Sets the maximum depth of individual trees.

Grid search with 3-fold cross-validation was used to evaluate combinations of these parameters, and the best configuration was merged with fixed parameters like `objective=’multiclass’` and `metric=’multi_logloss’`.

Model Training: The final LightGBM model was trained using the best hyperparameters identified during the tuning process, leveraging gradient-boosting decision trees for accurate multiclass classification.

Prediction and Evaluation: The model's predictions on the test set were evaluated using:

- **Accuracy:** Proportion of correctly classified samples.
- **Precision:** Proportion of true positives among predicted positives.
- **Recall:** Model's ability to capture all true positives.
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC:** Calculated using a one-vs-rest approach to measure performance across all classes.

Confusion Matrix: A confusion matrix was computed to analyze classification performance for each class, showing the distribution of true positives, false positives, false negatives, and true negatives.

ROC Curve: To visualize the model's performance, ROC curves were plotted for each class. The curves (Figure 9) demonstrate the trade-off between true positive and false positive rates. The area under each curve (AUC) was calculated to quantify the model's discrimination ability.

This systematic approach ensured the implementation of an optimized LightGBM model and comprehensive evaluation of its performance.

3) Convolutional Neural Network: A Convolutional Neural Network (CNN) was implemented for multiclass classification, leveraging its capability to hierarchically extract features. The methodology is as follows:

Data Preparation: Input data was reshaped into 3D format (`samples, features, 1`) for CNN processing. Labels were encoded as categorical for multiclass classification.

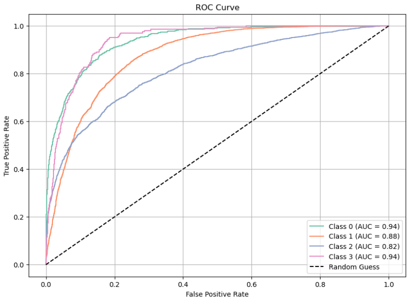

Fig. 9. ROC Curve for LightGBM

Model Architecture: The CNN model consisted of:

- **Convolutional Layers:** Two `Conv1D` layers with 32 and 64 filters, respectively, applied with ReLU activation to extract local patterns.
- **Pooling Layers:** `MaxPooling1D` layers reduced dimensionality while retaining essential features.
- **Flattening:** The feature maps were flattened into a 1D vector.
- **Fully Connected Layers:** Included a dense layer with 128 units (ReLU activation) and a Dropout layer to mitigate overfitting.
- **Output Layer:** The final dense layer used softmax activation for multiclass classification and sigmoid for binary tasks.

Model Training: The model was compiled using the Adam optimizer and trained for 10 epochs with a batch size of 32. Validation was performed using 20% of the training data.

Evaluation: The model was evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score. A classification report and confusion matrix were generated to provide detailed insights.

Receiver Operating Characteristic (ROC) Curve: ROC curves were plotted for each class, with the AUC values indicating performance. For binary classification, a single ROC curve was used. Figure 10 illustrates the ROC curves for this model.

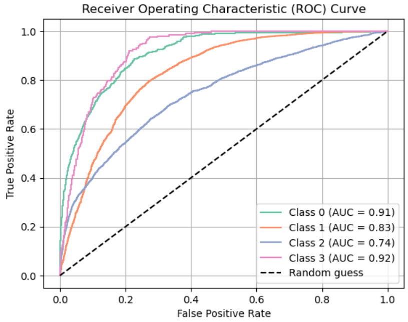

Fig. 10. ROC Curve for Convolutional Neural Network

This CNN-based approach effectively captured hierarchical features and provided robust classification performance across multiple classes.

4) Hybrid Model of Dual SVM and LightGBM: A hybrid model was constructed by combining predictions from multiple tuned classifiers: SVM with linear and RBF kernels, and LightGBM. The methodology is outlined below:

Hyperparameter Tuning: The SVM models were optimized for parameters `C`, `kernel`, and `gamma`, while the LightGBM model was tuned for `num_leaves`, `learning_rate`, and `max_depth`. The best configurations were determined using `GridSearchCV` with 3-fold cross-validation.

Hybrid Model Construction: The hybrid model combined predictions from the three classifiers using weighted averaging:

- **SVM (Linear):** Weight = 0.1
- **SVM (RBF):** Weight = 0.1
- **LightGBM:** Weight = 0.8

Evaluation Metrics: The hybrid model was evaluated using:

- **Accuracy:** Proportion of correctly classified samples.
- **Precision:** Proportion of true positives among all predicted positives.
- **Recall:** Ability to capture all true positives.
- **F1-Score:** Harmonic mean of precision and recall.

A detailed classification report and confusion matrix were generated to analyze performance.

Receiver Operating Characteristic (ROC) Curve: ROC curves were plotted for each class, with the AUC values indicating the classifier's ability to distinguish between classes. Figure 11 illustrates the ROC curves for the hybrid model.

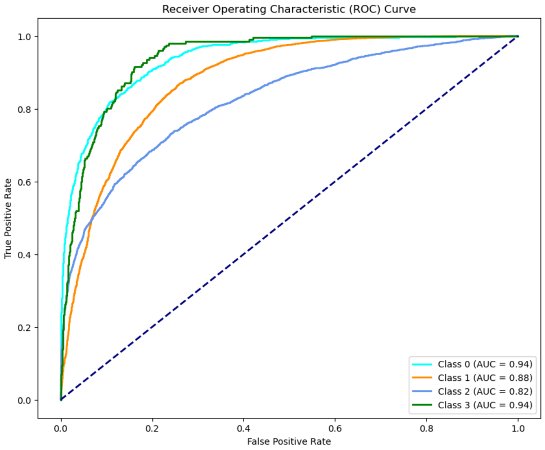

Fig. 11. ROC Curve for Hybrid - Dual SVM and LightGBM

The hybrid model effectively combines the strengths of individual classifiers to achieve robust performance.

V. Results

A. Dual Support Vector Machine

The dual SVM model, using both Linear and RBF kernels, demonstrates a balanced performance with an overall accuracy of 72.36%. The Linear SVM performs well for the class with the highest support (label 1), achieving a high recall (93%) and precision (75%). However, it struggles with classifying less frequent classes, such as label 3, where it shows a significantly lower recall of just 6%. The RBF SVM, with its optimal hyperparameters ('C': 10, 'gamma': 'scale'), also shows good performance for the majority class, but faces similar challenges with smaller classes. The confusion matrix highlights that the model is particularly challenged with label 3, which has fewer instances and poor classification results, as indicated by its low F1 score.

TABLE I
SVM Model Performance Metrics.

Metric	Score
Accuracy	0.7236
Precision	0.7113
Recall	0.7236
F1 Score	0.6986

TABLE II
Confusion matrix for the SVM model (0 = Normal, 1 = Atrial Fibrillation, 2 = Other, 3 = Noisy).

Predicted	True			
	0	1	2	3
0	326	213	274	2
1	31	5017	356	2
2	90	1385	1310	3
3	12	88	90	12

B. LightGBM

The LightGBM model demonstrates a solid performance with an accuracy of 75.95%, showcasing its ability to effectively classify the target variable. The precision and recall scores of 74.78% and 75.95%, respectively, indicate a balanced performance in both correctly identifying positive instances and minimizing false negatives. The F1 score of 74.48% further highlights the model's effectiveness in handling class imbalance. The confusion matrix reveals that the model performs well for the majority class (1), with fewer misclassifications in classes 0, 2, and 3. While the misclassifications for class 2 and class 3 are noticeable, the overall results suggest the model is reliable.

TABLE III
LightGBM Model Performance Metrics.

Metric	Score
Accuracy	0.7595
Precision	0.7478
Recall	0.7595
F1 Score	0.7448

C. Convolutional Neural Network

The performance of the Convolutional Neural Network (CNN) model is evaluated with an accuracy of 70.54%, showing reasonable predictive capability. The precision of 68.01% indicates the model's ability to correctly identify positive instances, while the recall of 70.54% shows it is

TABLE IV
CONFUSION MATRIX FOR THE LIGHTGBM MODEL (0 = NORMAL, 1 = ATRIAL FIBRILLATION, 2 = OTHER, 3 = NOISY).

Predicted	True			
	0	1	2	3
0	429	144	230	12
1	34	5022	331	19
2	122	1137	1504	25
3	19	77	65	41

effectively capturing relevant cases from each class. The F1 score of 66.78% balances precision and recall, suggesting moderate performance in classifying the target variables. The confusion matrix highlights some misclassifications, particularly in the predictions for classes 1 and 2, with class 0 being more accurately predicted. This suggests that while the model is performing well overall, there is room for improvement, especially in distinguishing between certain classes.

TABLE V
CNN MODEL PERFORMANCE METRICS.

Metric	Value
Accuracy	0.7054
Precision	0.6801
Recall	0.7054
F1 Score	0.6678

TABLE VI
CONFUSION MATRIX FOR THE CNN MODEL (0 = NORMAL, 1 = ATRIAL FIBRILLATION, 2 = OTHER, 3 = NOISY).

Actual	Predicted			
	0	1	2	3
0	279	246	289	1
1	15	5144	247	0
2	81	1633	1074	0
3	9	120	73	0

D. Hybrid Model of Dual SVM and LightGBM

The tuned metrics and confusion matrix showcase the effectiveness of the hybrid model combining dual SVM and LightGBM for classification. With an accuracy of 76.16%, the model demonstrates strong overall performance in distinguishing between the different labels, as indicated by its balanced precision (74.94%) and recall (76.16%). The F1 score of 74.71% reflects a good trade-off between precision and recall, ensuring minimal false positives and negatives. The confusion matrix highlights that the model is particularly proficient in classifying class '1', with fewer misclassifications in this category. However, there is still room for improvement in distinguishing class '2' and '3', where some misclassifications occur.

E. Model Performance Comparison

The following table summarizes the performance of the models based on accuracy, precision, recall, and F1 score:

TABLE VII
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1
Dual SVM(Linear+RBF)	0.7236	0.7113	0.7236	0.6986
LightGBM	0.7595	0.7478	0.7595	0.7448
CNN	0.7054	0.6801	0.7054	0.6678
Hybrid(SVM+LightGBM)	0.7615	0.7494	0.7615	0.7470

Model Comparison: The **Hybrid model** (SVM + LightGBM) outperforms the other models across all metrics, achieving an accuracy of 76.15%, along with the highest precision, recall, and F1 score. The **LightGBM** model also performs strongly with an accuracy of 75.95%, but the hybrid model slightly edges it out in performance. The **Dual SVM** model offers good recall but has a lower F1 score, while the **CNN** model shows relatively lower performance across all metrics.

VI. DISCUSSION

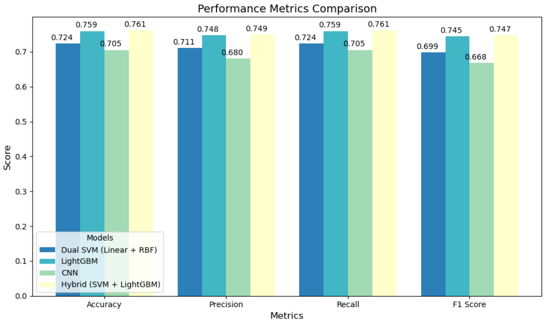

Fig. 12. Performance Metrics Comparison

The performance evaluation of various models highlights the strengths and weaknesses of each approach in classifying the target labels, particularly within imbalanced data scenarios. The Hybrid model, which combines Dual SVM and LightGBM, achieved the highest overall accuracy (76.16%), along with the best precision, recall, and F1 score. This indicates its superior ability to balance misclassifications across different classes. LightGBM, as an individual model, also demonstrated robust performance with an accuracy of 75.95%, showcasing its effectiveness in handling class imbalance. However, the Dual SVM, while excelling in recall for the majority class, struggled significantly with minority class predictions, as evidenced by its low F1 score and poor recall for the least frequent class. Similarly, the CNN model showed moderate performance with an accuracy of 70.54%, but its relatively lower F1 score suggested challenges in precision-recall tradeoffs, particularly for minority classes. These observations underscore the importance of model selection and hybridization to improve classification performance.

The confusion matrices further revealed critical insights into the classification capabilities of each model. Both LightGBM and the Hybrid model showed fewer misclassifications for the majority class (label 1), while still encountering difficulties with minority classes, particularly labels 2 and 3. The

CNN model, despite achieving a reasonable overall accuracy, struggled with consistent predictions for classes 2 and 3, which highlights the need for further optimization or additional techniques, such as data augmentation or cost-sensitive learning, to address imbalances. Meanwhile, the Hybrid model demonstrated its advantage by combining the strengths of Dual SVM and LightGBM, offering a more balanced performance across all metrics. This reinforces the value of integrating complementary models to enhance predictive accuracy and robustness in complex multi-class classification problems.

VII. Conclusion

In this study, the performance of four distinct models—Dual SVM, LightGBM, CNN, and a Hybrid model combining Dual SVM and LightGBM was evaluated on a challenging multi-class classification task. The Hybrid model emerged as the most effective approach, outperforming all other models across key performance metrics such as accuracy, precision, recall, and F1 score. Its ability to combine the complementary strengths of both SVM and LightGBM allowed for more accurate and balanced classification, especially when faced with class imbalances. LightGBM demonstrated strong performance on its own, confirming its suitability for handling complex classification tasks with imbalanced datasets, but the Hybrid model's integration of two diverse algorithms provided a notable edge in performance.

On the other hand, while the Dual SVM model showed solid recall for the majority class, its performance was hindered by its struggle with minority classes, reflecting the limitations of linear and RBF kernels in handling such imbalances. The CNN model, despite achieving reasonable accuracy, faced significant challenges in distinguishing between certain classes, particularly in terms of precision-recall balance. These findings underscore the importance of model selection and the potential of hybridization in enhancing the predictive capability of machine learning algorithms. The results also highlight the need for tailored approaches to address the complexities of multi-class classification, especially when class imbalance is a significant concern.

- **Potential Limitations:**
  Signal Length Variation: Sliding window approach with a fixed window size may miss important events in shorter ECG signals or not capture long-duration events that span multiple windows.
  Fixed Bandpass Filter: The 0.5–50 Hz filter may not be optimal for all ECG signals, potentially removing important components of the signal, especially in certain heart conditions.
  Feature Selection Bias: Extracted features may not capture the most relevant aspects of the ECG signal, potentially affecting model performance.
  Class Imbalance: Class imbalance is leading to potential bias toward the majority class and poor performance on the minority class.
- **Scope and Generalization:** The scope of this study is not limited to Atrial Fibrillation detection; the methods developed here could be generalized to other cardiac arrhythmia detection tasks using ECG signals. The Hybrid model's flexibility makes it suitable for various types of heart-related abnormalities, which can be valuable in clinical settings where accurate and timely detection is essential. Moreover, the ability to handle imbalanced data is a key feature of these models, particularly with ECG signals that often have skewed distributions, such as rare occurrences of AF. The generalization of the proposed models to other patient populations and ECG signal types should be tested to evaluate their robustness and applicability in real-world healthcare environments. The results of this study suggest that the Hybrid model exhibits strong generalization potential when applied to new, unseen ECG data. This is critical in real-world applications, where models must handle variations in data quality, patient demographics, and different sensor types. Future work could focus on assessing the model's performance across multiple datasets, including those with diverse patient cohorts, to ensure that the model does not overfit to the training data. Additionally, a broader generalization study could explore how these models can be deployed in mobile health devices for continuous AF monitoring, contributing to a more dynamic and scalable solution for real-time cardiac care.
- **Future Work:** Longitudinal Studies: Longitudinal ECG data could be used to examine how features evolve over time, potentially leading to more accurate predictions for chronic conditions. Multimodal Data: Additional patient data, such as heart rate and blood pressure, could be incorporated to enhance feature extraction and improve classification performance. Deep Learning Models: Deep learning models, such as CNNs or RNNs, could be explored to learn features directly from raw ECG data, which may improve classification performance. Personalized Models: Personalized models could be developed to adapt to individual patient characteristics, potentially offering more accurate predictions.

References

[1] Beau Bo-Sheng Chuang and Albert 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 Formative Research*, 8:e47803, 2024.

[2] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[3] Gamal G. N. Geweid and Jiande D. Z. Chen. Automatic classification of atrial fibrillation from short single-lead ecg recordings using a hybrid approach of dual support vector machine. *Expert Systems with Applications*, 198:116848, 2022.

[4] G. Hindricks, T. Potpara, N. Dagres, E. Arbelo, J. J. Bax, C. Blomström-Lundqvist, G. Boriani, M. Castella, G. A. Dan, P. E. Dilaveris, et al. 2020 esc guidelines for the diagnosis and management of atrial fibrillation. *European Heart Journal*, 42(5):373–498, 2021.

[5] P. Langley, E. J. Bowers, and A. Murray. Electrocardiogram techniques for the detection of atrial fibrillation. *Journal of the American College of Cardiology*, 48(4):994–1001, 2006.

[6] Mou Wang, Sylwan Rahardja, Pasi Fränti, and Susanto Rahardja. Single-lead ecg recordings modeling for end-to-end recognition of atrial fibrillation with dual-path rnn. *Biomedical Signal Processing and Control*, 79:104067, 2023.

[7] P. A. Wolf, R. D. Abbott, and W. B. Kannel. Impact of atrial fibrillation on mortality, stroke, and medical costs. *Archives of Internal Medicine*, 158(3):229–234, 1998.