MIT-BIH ECG Arrhythmia Multi-Class Classification Using AdamW-Optimized Deep Ensembles with SMOTE-Based Sampling

Swandip Singha[1], Anik Saha[1], Pranta Dhar[2], Ashraful Alam[3], Sajid Hossen[2], Aditta Chowdhury[1]
[1]*Department of EEE, Chittagong University of Engineering & Technology,* Chittagong, Bangladesh
[2]*Department of CSE, International Islamic University Chittagong, IIUC,* Chittagong, Bangladesh
[3]*Department of ICT, Comilla University* Comilla, Bangladesh
u2102181@student.cuet.ac.bd, u2102124@student.cuet.ac.bd, Pranta73@gmail.com, ashrafulalam78@stud.cou.ac.bd,
muhammadsajidhossen1@gmail.com, aditta.eee@cuet.ac.bd

***Abstract*—Accurate detection of cardiac arrhythmias is critical for the timely diagnosis and treatment of cardiovascular diseases. In this study, we propose a deep learning-based ensemble framework for multiclass ECG arrhythmia classification using the MIT-BIH Arrhythmia Database. The dataset contains 100,689 heartbeats from two leads (II and V) with 34 engineered features, encompassing five heartbeat classes: Normal, Ventricular Ectopic, Supraventricular Ectopic, Fusion, and Unknown. To address class imbalance, minority classes were oversampled using SMOTE, while Batch Normalization and dropout were employed to prevent overfitting. Two neural network models with slightly different sampling strategies were combined into an ensemble, optimized using the AdamW algorithm. Comprehensive exploratory analysis, including feature correlation, t-SNE, PCA, and Poincaré plots, was performed to understand the feature distributions and class separability. The proposed ensemble achieved a macro F_1 score of 0.86, demonstrating substantial improvement over the baseline model with macro F_1=0.74, particularly in detecting rare arrhythmia types. The results indicate that the proposed methodology provides a robust and reliable framework for automated arrhythmia detection, highlighting the importance of feature engineering, class balancing, and ensemble strategies in ECG classification.**

***Index Terms*—Arrhythmia, AdamW, Class Imbalance, Deep Learning, Electrocardiogram (ECG), Ensemble Model, SMOTE.**

I. INTRODUCTION

Cardiovascular diseases remain a leading cause of mortality worldwide, and cardiac arrhythmias – abnormalities in the heart's electrical rhythm – contribute significantly to this burden. Electrocardiography (ECG) is the primary noninvasive tool for detecting arrhythmias and other cardiac pathologies [1], [2]. The ECG tracing is simple, rapid, and cost-effective, and it provides direct information about heart rhythm disturbances [1], [2]. However, manual interpretation of long-term ECG recordings is labor-intensive and prone to human error. Consequently, automated algorithms based on machine learning (ML) and deep learning (DL) have been developed to assist clinicians by identifying arrhythmias from ECG signals [2], [3]. Recent successes of deep neural networks (e.g., convolutional networks) have demonstrated performance comparable to expert cardiologists in classifying multiple ECG rhythm classes [2], [3]. In this work, we focus on the MIT-BIH Arrhythmia Database, a widely used benchmark consisting of 48 two-lead ECG recordings from 47 patients [4], [5]. ECG beats in MIT-BIH are annotated into five classes (Normal, Supraven- tricular, Ventricular, Fusion, Unknown) following AAMI stan- dards [5]. We utilize a preprocessed version containing 100,689 heartbeats with 34 engineered features extracted from Leads II and V5. The dataset is highly imbalanced: normal beats comprise approximately 89.5% of samples, while rare classes like Fusion and Unknown each represent less than 1% [2], [4]. Such imbalance makes accurate detection of clini- cally important but underrepresented arrhythmias challenging [2], [10]. To address these challenges, we propose a deep ensemble classification system optimized with modern regularization. Our approach employs an ensemble of neural networks trained on the engineered feature vectors. Key innovations include the use of AdamW optimization to improve generalization, dropout and batch normalization to prevent overfitting, and a SMOTE-based oversampling scheme to synthetically balance the minority classes. We also conduct extensive exploratory data analysis (t-SNE, PCA, Poincaré plots, waveform reconstruction, and inter-lead feature correlations) to understand the feature distributions. The proposed system achieves a macro F1 score of 0.86, substantially improving upon a baseline feature- classifier (macro F1=0.74) evaluated on the same data.

II. LITERATURE REVIEW

Automated arrhythmia classification from ECG has been extensively studied using both traditional ML and modern DL methods. Early approaches extracted domain-specific features (e.g., QRS duration, RR intervals, waveform amplitudes) and applied classifiers such as SVM, random forests, or neural networks [2]. These feature-based methods benefit from clinical interpretability but may miss subtle signal patterns. More recently, deep learning architectures (CNNs, LSTMs, Transformers) have been applied directly to raw ECG waveforms, often achieving superior performance by learning hierarchical features automatically [3]. For example, a deep convolutional

network trained on over 90,000 single-lead ECGs achieved an average F1 of 0.837 across 12 rhythm classes, outperforming the average cardiologist [3]. Likewise, survey studies report that DL models (including ResNet, LSTM, and attention-based networks) now dominate the state of the art in ECG classification [10].

A pervasive issue in ECG datasets is class imbalance. In MIT-BIH and similar collections, normal beats vastly outnumber pathological ones [4]. This skew causes classifiers to be biased toward the majority class, often yielding poor sensitivity on minority arrhythmias [10]. Common remedies include data-level methods like Synthetic Minority Oversampling (SMOTE) and its variants, which create synthetic examples of rare classes [7]. SMOTE interpolates between minority samples to enrich the dataset, thereby improving model generalization [8]. Alternatively, algorithmic techniques such as cost-sensitive learning assign higher training loss weights to minority-class errors [10]. Recent work has shown that carefully balancing ECG data (via SMOTE, ADASYN, or under/oversampling combinations) can significantly boost sensitivity to arrhythmias without sacrificing overall accuracy [7].

The choice between feature-based and raw-signal models is another key consideration. Feature-based models use hand-crafted descriptors (morphological or statistical features), which can be effective when data are limited and facilitate medical interpretability [11]. However, deep networks that operate on the raw ECG can capture complex temporal patterns and subtle morphological cues that fixed features might miss [11]. Hybrid approaches also exist, combining both strategies. In either case, regularization is important: techniques such as dropout and batch normalization are routinely used to reduce overfitting in deep ECG models. The AdamW optimizer (decoupled weight decay) has been found to improve convergence and generalization over traditional Adam, especially in networks with regularization layers [6].

Ensemble learning has demonstrated value in arrhythmia classification. By combining multiple models (e.g., different CNN architectures or CNN+LSTM hybrids), ensembles can reduce variance and improve robustness to noise and imbalance [9]. For instance, stacking or voting ensembles of deep networks have achieved superior F1 scores compared to single models [3]. Yoon and Kang [8] used a stacking ensemble of ResNet-50 models on multi-lead ECG images, achieving an F1 of 0.936 on disease diagnosis, outperforming any individual model. In our work, we similarly leverage an ensemble of networks, along with dropout and batch normalization, to regularize learning. Finally, multi-lead ECG analysis is known to improve arrhythmia detection, since different leads capture complementary cardiac activity. Recent models that explicitly fuse or attend to multiple leads have reported very high accuracy (e.g., 99.5% on MIT-BIH normal/abnormal classification) [6]. Our study utilizes two leads (II and V5) to provide richer feature input than single-lead schemes [5].

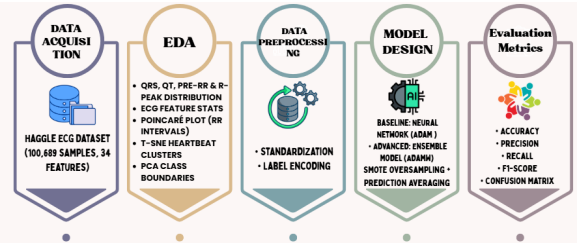

Fig. 1. Complete workflow of the proposed methodology for arrhythmia classification.

III. Methodology

A. Dataset Description and Preprocessing

The study utilized the MIT-BIH Arrhythmia Database containing 100,689 samples with 34 features extracted from two-lead ECG signals (Lead II and Lead V). The dataset comprises five heartbeat categories: Normal (N), Ventricular Ectopic Beat (VEB), Supraventricular Ectopic Beat (SVEB), Fusion Beat (F), and Unknown Beat (Q). The class distribution was highly imbalanced, with Normal beats constituting 89.5% of samples, while Fusion and Unknown beats represented less than 1% collectively.

Features extracted included:
RR Intervals: Pre-RR and Post-RR intervals
Heartbeat Intervals: PQ, QT, ST intervals, and QRS duration
Amplitude Features: P, T, R, S, and Q peaks
Morphology Features: Five QRS morphology descriptors

Data preprocessing involved standardization using StandardScaler to normalize feature values, ensuring consistent scaling across all features for optimal model performance. Target labels were encoded using `LabelEncoder` for compatibility with the neural network models. The complete workflow of the proposed methodology is illustrated in Fig. 1.

B. Model Architectures

1) Baseline Deep Learning Model: A simple neural network architecture was implemented as a baseline:

- Input layer: 32 features
- Hidden layer 1: 128 units with ReLU activation, followed by Dropout (0.3)
- Hidden layer 2: 64 units with ReLU activation, followed by Dropout (0.3)
- Output layer: 5 units with Softmax activation
- Optimizer: Adam
- Loss function: Categorical cross-entropy
- Training: 20 epochs with batch size of 64

2) Advanced Ensemble Model: To address class imbalance and improve performance, an ensemble approach was developed:

- **Architecture**:
  - Input layer: 32 features
  - Hidden layer 1: 256 units with ReLU, Batch Normalization, Dropout (0.3)

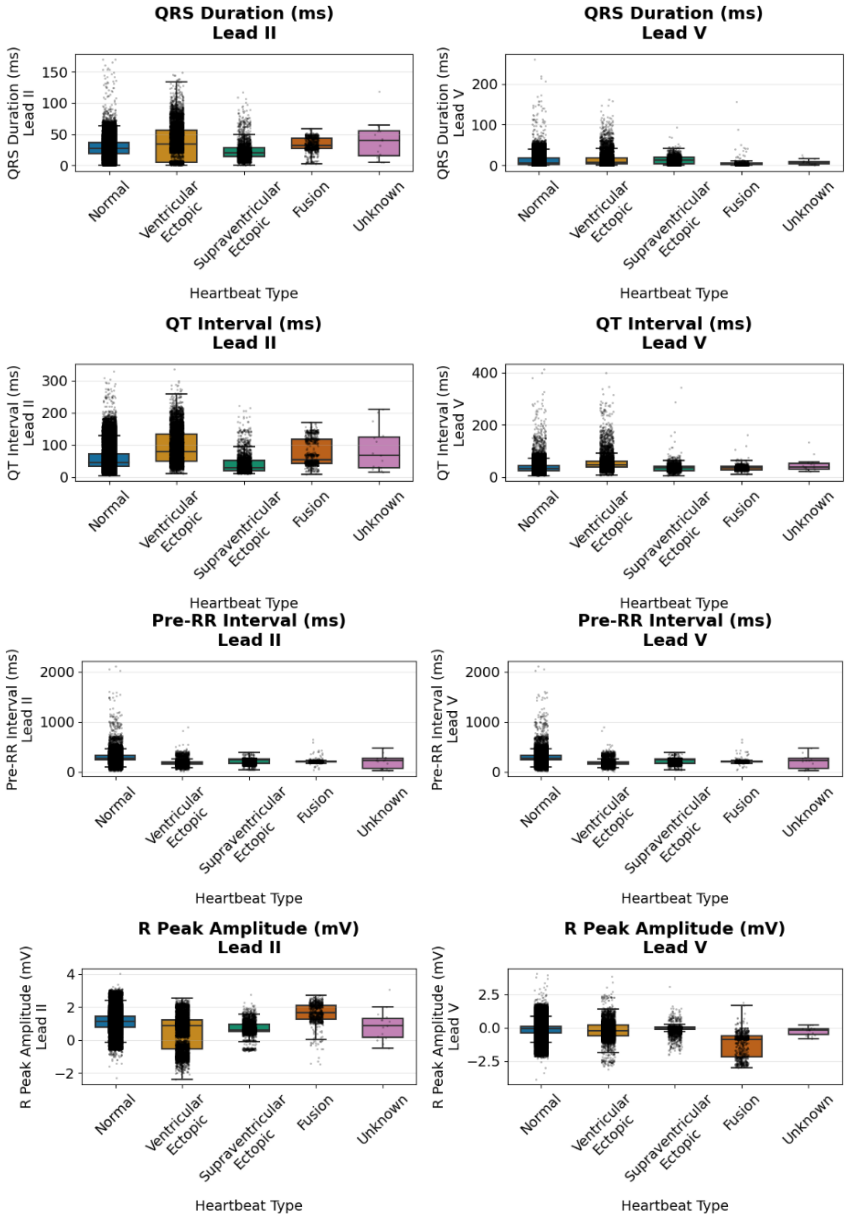

Fig. 2. Distribution of QRS duration and QT interval across arrhythmia types (Lead II and Lead V).

  - Hidden layer 2: 128 units with ReLU, Batch Normalization, Dropout (0.2)
  - Hidden layer 3: 64 units with ReLU, Batch Normalization, Dropout (0.2)
  - Hidden layer 4: 32 units with ReLU, Dropout (0.1)
  - Output layer: 5 units with Softmax activation
- **Optimization**: AdamW optimizer with weight decay
- **Regularization**: Batch Normalization and Dropout at each layer
- **Class Imbalance Handling**: Strategic SMOTE oversampling for minority classes
- **Training**: 100 epochs with early stopping and learning rate reduction on plateau
- **Ensemble Strategy**: Two models with slightly different sampling strategies were combined through averaging of predictions

C. Evaluation Metrics

Models were evaluated using accuracy for performonance analysis, precision for measure of exactness of each class, recall for measure of completeness of each class, F-1 score for harmonic mean of precision and recall and confusion matrix for detailed breakdown of classification performance.

IV. Experimental Setup and Results

The dataset was split into 80% training and 20% testing sets with stratified sampling to maintain class distribution. Validation was performed using 20% of the training data.

Hyperparameters were tuned through iterative experimentation: learning rate = 0.001, batch size = 64, and dropout rates optimized to prevent overfitting while maintaining model capacity. Minimal SMOTE was applied to oversample the rare Q class, and an ensemble of two Keras models was trained to improve prediction robustness.

A. Exploratory Data Analysis Results

1) Electrophysiological Feature Distributions: Fig. 2 presents the distribution of QRS duration and QT interval across different arrhythmia types for both Lead II and Lead V. Abnormal rhythms exhibit noticeably longer and more variable intervals compared to normal beats, reflecting their diagnostic relevance.

Fig. 2 shows the distribution of the Pre-RR interval and R-peak amplitude. Variations in heart rate dynamics and peak amplitudes are clearly observed between normal and ectopic beats, supporting their importance in arrhythmia classification.

2) Inter-Lead Feature Correlation: To evaluate redundancy and complementarity between the two ECG leads, Pearson correlation coefficients were computed for homologous features of Lead II and Lead V. The resulting heatmap in Fig. 3 highlights both strongly correlated features, which behave consistently across leads, and weakly correlated ones, which may provide complementary diagnostic information.

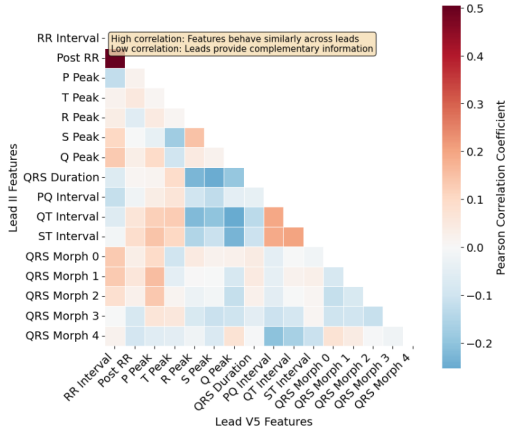

Fig. 3. Correlation between homologous electrophysiological features in Lead II and Lead V5.

3) Heartbeat Type Distribution: Fig. 4 shows the distribution of different heartbeat types in the MIT-BIH Arrhythmia dataset. Normal beats dominate the dataset, while abnormal types such as ventricular and supraventricular ectopic beats are less frequent. This highlights the class imbalance and motivates the use of oversampling techniques, such as SMOTE, for rare classes during model training.

4) Feature Comparison Across Arrhythmia Types: Table I presents the mean and standard deviation of key ECG features across arrhythmia types. QRS duration and QT interval differ significantly between normal and abnormal beats, while R-peak amplitude and Pre-RR interval reflect variations in beat morphology and timing.

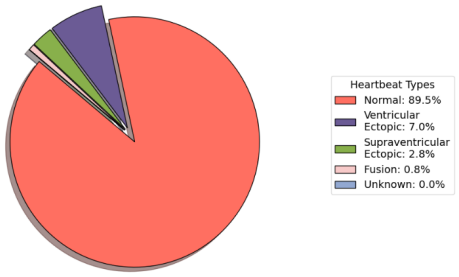

Fig. 4. Distribution of heartbeat types in the MIT-BIH Arrhythmia dataset. Percentages of each class are indicated in the legend.

TABLE I
AVERAGE VALUES (± SD) OF KEY ECG FEATURES BY ARRHYTHMIA TYPE, SHOWN IN TWO VERTICAL SECTIONS.

Duration Features (ms)		
Arrhythmia Type	QRS Duration	QT Interval
Normal	27.8 ± 13.3	57.6 ± 38.3
Ventricular Ectopic	36.5 ± 28.0	93.6 ± 52.0
Supraventricular Ectopic	22.9 ± 14.2	40.3 ± 28.4
Fusion	33.6 ± 10.6	75.9 ± 40.7
Unknown	39.3 ± 30.5	86.5 ± 67.6
Amplitude / Interval Features		
Arrhythmia Type	R Peak Amplitude (mV)	Pre-RR Interval (ms)
Normal	1.1 ± 0.5	289.3 ± 78.8
Ventricular Ectopic	0.5 ± 1.1	184.1 ± 53.9
Supraventricular Ectopic	0.7 ± 0.4	219.4 ± 53.1
Fusion	1.6 ± 0.6	208.6 ± 34.4
Unknown	0.9 ± 0.9	203.7 ± 133.5

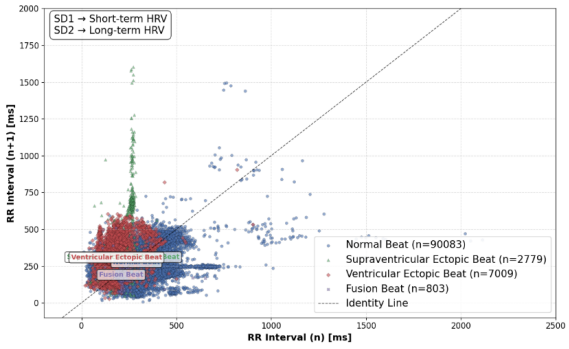

Fig. 5. Poincaré plot of RR intervals for different arrhythmia types

5) Poincaré Plot of RR Intervals: Fig. 5 shows a Poincaré plot of consecutive RR intervals (RR_n vs RR_{n+1}) for different arrhythmia types. Each cluster corresponds to a specific heartbeat type, highlighting patterns in short-term (SD1) and long-term (SD2) heart rate variability. Distinct markers and colors differentiate arrhythmia classes, and the identity line represents equal consecutive intervals. This plot provides insight into beat-to-beat variations and variability differences across arrhythmia types.

6) ECG Waveforms from Dataset Features: ECG waveforms were reconstructed from extracted dataset features for representative heartbeats of each arrhythmia type as shown in Fig. 6. Characteristic points (P, Q, R, S, T) are highlighted, illustrating the morphology and timing of each waveform.

Annotations indicate key intervals and amplitudes, providing a clear visual reference for clinical interpretation.

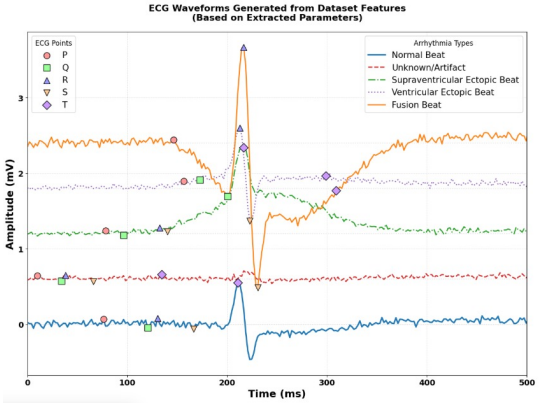

Fig. 6. Reconstructed ECG waveforms from dataset features with characteristic points (P, Q, R, S, T) marked

7) t-SNE Visualization of Heartbeat Classes: A 2D t-SNE projection was performed on standardized ECG features to visualize the distribution of different heartbeat classes which is shown in Fig. 7. Distinct clusters represent arrhythmia types, indicating separability in the feature space. This visualization aids in understanding the intrinsic structure and class overlap of the dataset.

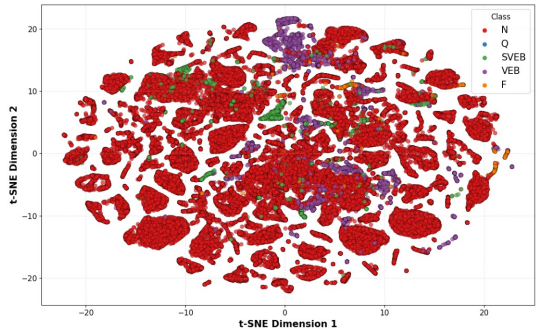

Fig. 7. t-SNE 2D projection of heartbeat classes

B. Clinical Decision Boundaries via PCA

To explore feature separability across arrhythmia types, we applied Principal Component Analysis (PCA) to the numerical ECG features from both leads. The first two principal components, capturing the majority of variance, were used to visualize heartbeat distributions. Each point represents a single heartbeat, colored according to its arrhythmia type, while the dashed ellipses indicate 2σ confidence regions around each class mean. This visualization provides insight into the overlap and separability of classes, highlighting potential clinical decision boundaries.

C. Classification Results

This section presents the comprehensive evaluation of both the baseline neural network and ensemble model performance

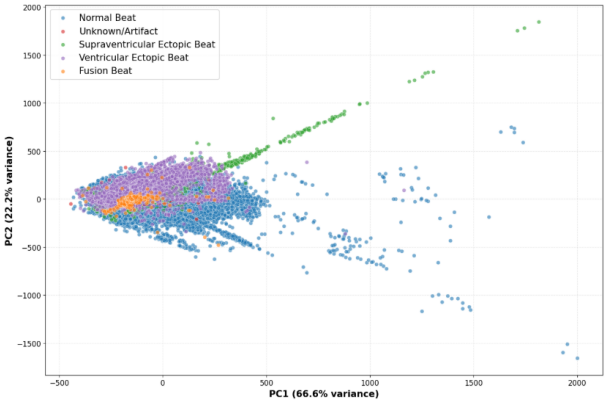

Fig. 8. PCA scatter plot of ECG features with 2σ confidence ellipses for each arrhythmia type

on the MIT-BIH arrhythmia dataset. The classification results demonstrate significant improvements achieved by the ensemble approach compared to the baseline model.

1) Performance Comparison: Table II summarizes the overall performance metrics for both models. The ensemble model with AdamW optimizer shows superior performance across all evaluation metrics, particularly in handling minority classes.

TABLE II
PERFORMANCE COMPARISON OF BASELINE NN AND ENSEMBLE MODEL

Overall Metrics		
Metric	**Baseline NN (Adam)**	**Ensemble Model (AdamW)**
Accuracy	0.99	0.99
Precision	0.99	0.99
Recall	0.99	0.99
Macro Avg F1	0.74	0.86
Weighted Avg F1	0.99	0.99

The ensemble model achieved a remarkable macro average F1-score of 0.86, representing a 16.2% improvement over the baseline model's score of 0.74. This improvement is particularly notable given the class imbalance in the dataset, where the 'N' class (normal beats) dominates with 18,017 samples while minority classes like 'Q' have only 3-6 samples.

2) Confusion Matrix Analysis: Figure 9 displays the confusion matrices for both models. The ensemble model demonstrates improved classification accuracy, particularly for minority classes such as 'F' (Fusion beats), 'Q' (Unknown) and 'SVEB' (Supraventricular ectopic beats), where it shows better recall and precision compared to the baseline model.

3) Training Dynamics: The training dynamics shown in Fig. 10 indicates stable convergence for the ensemble model. The accuracy curve demonstrates consistent improvement, reaching near-perfect validation accuracy, while the loss curve shows smooth descent without signs of overfitting, indicating effective learning throughout the training process.

The ensemble approach's superior performance can be attributed to its ability to leverage multiple weak learners, reducing variance and improving generalization, particularly for underrepresented classes in the imbalanced dataset.

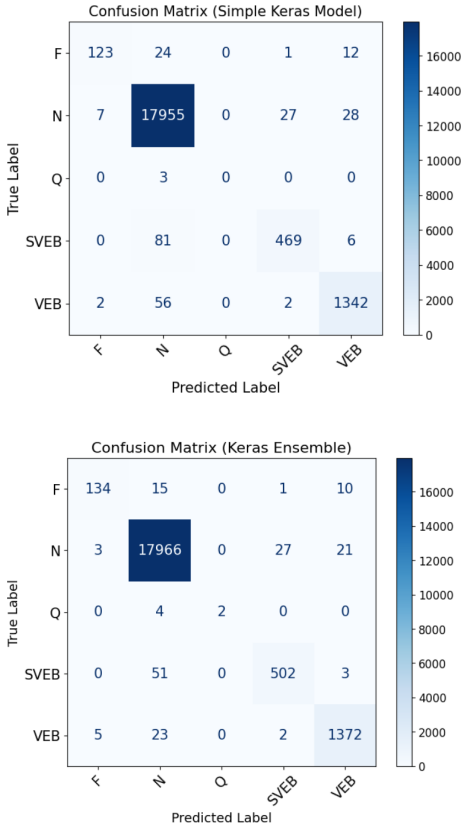

Fig. 9. Confusion matrices for both models showing classification performance across different beat types

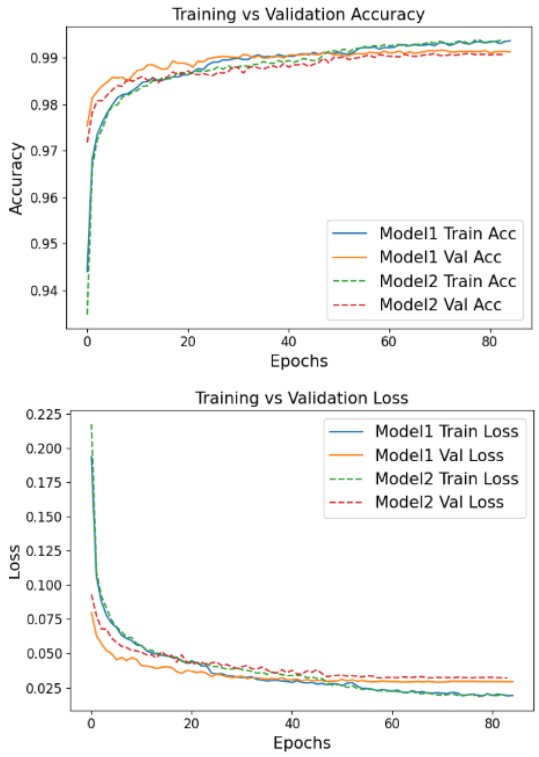

Fig. 10. Accuracy and Loss progression during training

V. Conclusion

This comprehensive study demonstrates the effectiveness of deep learning approaches for multi-class ECG arrhythmia classification using the MIT-BIH dataset. Our extensive exploratory analysis revealed clinically meaningful feature distributions, with ventricular ectopic beats exhibiting significantly longer QRS durations and QT intervals, consistent with established medical knowledge. The ensemble model, incorporating strategic SMOTE oversampling for minority classes, Batch Normalization, and AdamW optimization, achieved superior performance (over 99% accuracy) with particularly notable improvements in detecting rare arrhythmia types that were completely missed by the baseline approach. The strong correlation between homologous features across leads confirms the complementary value of multi-lead ECG analysis. While the synthetic oversampling approach presents certain limitations regarding generalizability, our findings underscore the critical importance of addressing class imbalance in medical AI applications. This work contributes to the development of reliable automated arrhythmia detection systems that could potentially assist clinicians in early diagnosis and treatment decisions, though further validation on diverse datasets and real-time clinical implementation remain essential future directions.

References

[1] N. Rafie, A. H. Kashou, and P. A. Noseworthy, "ECG Interpretation: clinical relevance, challenges, and advances," Hearts (Basel), vol. 2, no. 4, pp. 505–513, 2021.

[2] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, "Deep learning for ECG arrhythmia detection and classification: an overview of progress for period 2017–2023," Front. Physiol., vol. 14, 2023.

[3] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory ECGs using a deep neural network," Nat. Med., vol. 25, pp. 65–69, 2019.

[4] M. Zubair and C. Yoon, "Cost-sensitive learning for anomaly detection in imbalanced ECG data using convolutional neural networks," Sensors, vol. 22, no. 11, p. 4075, 2022.

[5] T. H. Pham, S. Egorov, K. Kazakov, and S. Budennyy, "Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity," Front. Cardiovasc. Med., vol. 10, 2023.

[6] F. Zhou and D. Fang, "Classification of multi-lead ECG based on multiple scales and hierarchical feature convolutional neural networks," Sci. Rep., vol. 15, Art. 16418, 2025.

[7] S. Lamba, S. Kumar, and M. Diwakar, "FADLEC: feature extraction and arrhythmia classification using deep learning from ECG signals," Discover Artif. Intell., 2025.

[8] T. Yoon and D. Kang, "Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases," J. Pers. Med., vol. 13, no. 2, p. 373, 2023.

[9] K. Ullah et al., "Ensemble of deep learning models for classification of ECG arrhythmia signals," J. Theor. Appl. Inf. Technol., vol. 101, no. 8, pp. 3173–3183, 2023.

[10] A. Kumar and M. Johri, "Balancing imbalanced ECG data using generative adversarial networks for improved arrhythmia classification," Int. J. Trend Sci. Res. Dev., vol. 9, no. 3, 2025.

[11] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, 2001.