Enhancing Processing Unit for Diagnosis Information With Deep Learning Models in ECG Device

Intissar Zaway[1], Lassaad Zaway[2], and Kaouthar Mansour[3]

[1] CEM-Laboratory, National School of Engineering of Sfax, University of Sfax, Tunisia,

[2] CES-Laboratory, National School of Engineering of Sfax, University of Sfax, Tunisia,

[3] IResCoMath-Laboratory, National School of Engineering of Gabes, University of Gabes , Tunisia,

Abstract

Developing clinical decision requires an accurate and robust categorization of physiological data, especially in cardiovascular monitoring.
However, some aberrations frequently taint these real-world signals, making IA models should be extremely robust in their learning process.
In this work, we tackle the challenge of developing models that are resistant in the presence of deteriorated signals in addition to achieving good classification performance.
Incorporating ECG signals and physiological features, a comparison study of IA (including Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), Bidirectional LSTM (BiLSTM), Random Forest (RF), and Support Vector Machine (SVM)) performance on classification anomalies (as Arrhythmia, Fibrillation, Tachycardia, Bradycardia) is primarily developed. Our contribution is to offer a thorough benchmark that evaluates the model's accuracy and resilience to external disturbances (such as noise, drift, missing data, heart rate variability, and motion abnormalities) .
By the trial findings, the CNN model shows the best capacity to accurately detect true positives and achieves the greatest accuracy (99.33%) and recall (98.40%). The LSTM model, on the other hand, performs nether, with a recall of 86.56% and an accuracy of 43.76%.
Each DL model's (CNN, LSTM and BiLSTM) resilience to different data problems is then examined. In the majority of cases, CNN routinely performs better than other models, with BiLSTM coming in second. It implements the greatest stability score (0.796) according to the comparative robustness analysis, demonstrating its dependability and suitability for use in this practical application.

Keywords: Classification - ECG signals - Machine Learning - Deep Learning - Robustness

1 Introduction

The World Health Organization estimates that cardiovascular diseases (CVDs) claim 17.9 million lives annually, making them the world's top cause of death [1]. The electrocardiogram (ECG) is still the most widely available, reasonably priced, and non-invasive diagnostic technique for determining the electrical activity of the heart [2]. Arrhythmia, ischemia, myocardial infractions, and conduction abnormalities are just a few of the cardiac abnormalities that can be identified with the help of ECG signals [3].
The need for automated and intelligent systems that can process and interpret ECG signals in real time is rising as wearable sensors, portable monitoring devices, and large-scale ECG databases become more widely available [4][5]. Even while manual analysis is accurate in clinical settings, it is time-consuming, prone to inter and intra observer variability, and challenging to scale, particularly in emergency care or telemedicine settings[6].
In this regard, Machine Learning (ML) [7] and Deep Learning (DL) [8] methods have become successful techniques for classifying ECGs and identifying anomalies [9].
Related worK:
When trained on carefully crafted features taken from ECG signals, such as time-domain, frequency-domain, and morphological traits, machine learning models have shown excellent performance.

Support Vector Machines (SVM) is employed in [10][11] to classify the anomalies in ECG signals. Results prove the efficacy of the proposed algorithm with 0.98 of accuracy.
Authors in [12] use k-Nearest Neighbors (k-NN) for the prediction of ECG classes and findings show the effectiveness of this methodlogy in categorizing of the ECG signals.
Random Forests (RF) in [13] improve their performance in detecting the classes of ECG signals. In the same application, Gradient Boosting Machines (GBM) in [14] have shown an excellent performance.
In recent years, DL models like Transformer-based architectures, Long Short-Term Memory networks (LSTMs), and Convolutional Neural Networks (CNNs) in [15] have demonstrated the capacity to build robust, hierarchical representations straight from raw data, obviating the necessity for human feature extraction.
In order to determine the best techniques for classifying ECG rhythms, this study evaluates the performance of deep learning architectures (CNN, BiLSTM, LSTM) versus more conventional machine learning techniques (Random Forest, SVM).
Despite the success of these models in controlled research environments, several critical challenges remain that hinder their deployment in real-world clinical applications.
In other hand, Static and clean datasets under optimal conditions are used to train and assess a large number of published models. The diversity observed in clinical practice, such as different patient demographics, comorbidities, and signal capture configurations, is frequently missing from these datasets[16]. Consequently, models can fail to apply to new, unknown patients and instead overfit to characteristics unique to the dataset.
Actual world Noise, baseline drift, lead disconnections, and artifacts (such as motion or muscle noise) commonly impair ECG recordings[17]. Patient safety in automated diagnostic systems depends on how models react to such perturbations, which is not sufficiently tested in the majority of current investigations.
Furthermore, common rhythms like normal sinus rhythm greatly outnumber uncommon but serious abnormalities like ventricular fibrillation or premature ventricular contractions in clinical datasets, which frequently show a high degree of class imbalance. The most clinically significant events may go undetected by models trained on such unbalanced data, yet overall accuracy may be excellent[18].
Several works have tried to tackle some of these issues. CNN-based models have achieved accuracies over 95% in the Arrhythmia Database, which has become a common benchmark for arrhythmia detection. Better temporal modeling skills have been demonstrated by hybrid models that combine CNN and LSTM. To improve feature discrimination, Data augmentation technics designs have also been suggested [19][20].
To increase robustness, some researchers investigate data augmentation or adversarial training. Systematic assessments under domain shifts, in inter-patient settings, and at different noise levels are still very rare, nevertheless.
By putting out a thorough and exacting framework for ECG classification and anomaly detection utilizing both ML and DL models, this work seeks to close the gap between experimental success and clinical usability. Our primary contributions are:
Data Augmentation and Regularization: In order to overcome class imbalance and enhance generalization, we investigate and contrast the efficacy of methods including mixup, SMOTE, and synthetic ECG generation (using GANs).
Multi-Model Benchmarking: Using a variety of publically accessible ECG datasets, we develop and assess a suite of machine learning and deep learning models, from sophisticated CNN, LSTM, RF, BiLSTM, and SVM utilizing statistical and morphological features.
Robustness Evaluation: To evaluate the resilience of each model under different perturbation conditions, such as baseline Drift, signal with missing data, motion artifact, Heart rate variation and additive noise (Gaussian, powerline interference), we conduct controlled experiments. We use metrics such as the robustness index, F1-score, and AUC to quantify performance decline.
Inter-Patient Generalization: Using inter-patient splits instead of random cross-validation, which more closely resembles real deployment settings, we examine how well models trained on a subset of patients perform on completely unseen individuals.
To conclude, this work attempts to contribute to the development of ECG analysis systems that are not only accurate but also reliable, trustworthy, and appropriate for real-world healthcare applications by fusing methodological rigor with pragmatic concerns.
"The paper is organized as follows: The process is described in general in Section 2. The preparation and collection of data using the electrocardiograph equipment are the main topics of Section 3. The preprocessing module and feature extraction are described in depth in Section 4. The training algorithms for

deep learning and machine learning are presented in Section 5. The perturbation generating mechanism is presented in Section 6. Lastly, the findings and discussions are shown in Section 7.

2 Methodology Description

Figure 1 illustrates the flowchart of the methodology. To begin with the Data Simulation which includes all the process of data acquisition and preparation, followed by Exploratory Data Analysis (EDA) Visualization.
After that, the process splits into two parallel tracks: Feature Extraction, which leads to RF/SVM training, and Spatial Features, which leads to CNN/LSTM/BiLSTM neural network training. Model Evaluation is the next step where the two algorithms arrive, and from there they diverge into Metrics Curves and Model Comparison before ending at Insights Reporting.

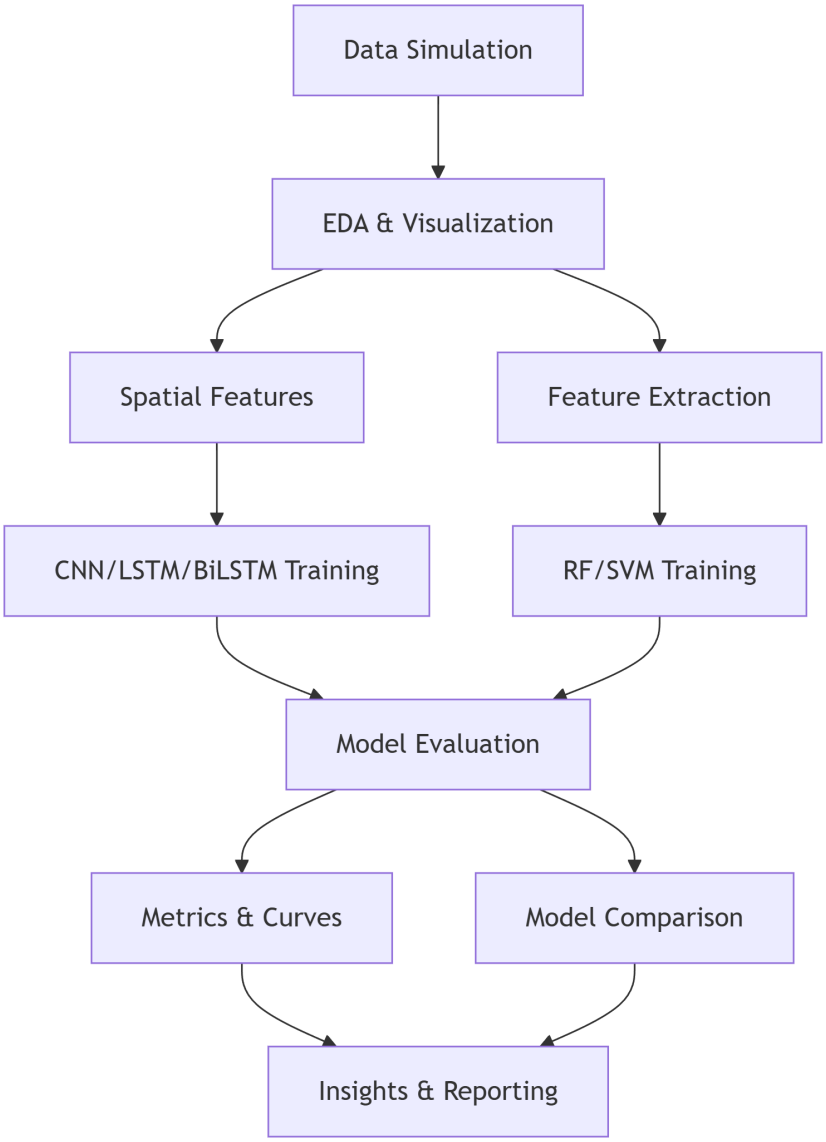

Figure 1: Methodlogy Flowchart

3 Data Acquisition and Preparation: COMEN Electrocardiograph Device

The dataset utilized in this study comprises multivariate time-series data derived from electrocardiogram (ECG) signals collected involving 150 patients. Each session includes continuous ECG recordings captured over a typical observation period (file.D25 type).
Data acquisition was conducted and cleaned using a stratified sampling strategy to ensure balanced representation across relevant classes and anomalies. The dataset originates from real-world clinical environments within a university hospital in Tunisia. To preserve patient confidentiality and align with ethical guidelines concerning sensitive health information, specific institutional identifiers have been intentionally omitted due to the limited cohort size and potential re-identification risks.
The authors extend their sincere appreciation to the hospital staff for their valuable support throughout the data collection and research activities.
The preprocessing steps were applied based on manufacturer industry. The filtering follows these main steps:
DFT Filter: used at 0.05 Hz to eliminate motion artifacts and extremely low frequency baseline drift while maintaining the integrity of important waveform elements, especially the ST segment.
AC Filter: permuts power line interference suppression. The frequency is set at 50 Hz for 220V systems.
EMG Filter: frequently turned off during acquisition, unless there was a severe muscular artifact. This method reduces the excessive attenuation of the ECG signal's with higher-frequency components.
Lowpass Filter: set to 100 Hz, it attenuates high frequency noise while permitting the transmission of therapeutically significant signal information.
The resultant ECG signals were minimally affected and appropriate for further feature extraction and diagnostic analysis thanks to this preprocessing technique.
To obtain the ECG waveform representing all 12 leads, equation 1 is employed:

$$\mathrm{ECG_t}(t) = \sqrt{\frac{1}{12}\sum_{i=1}^{12}\mathrm{ECG}_i(t)^2} \tag{1}$$

For the process of data augmentation and an equal class distribution, a basic duplication is used and on signals, a small random transformation is also applied. This process has the utility of obtain 150 ECG samples well structured and equal class distribution (each class has 30 samples).
This work offers a comparative study of several deep learning and machine learning models for automatic ECG signal classification. General observation of the Five selected categories are tabulated in 1 for the model designing and training.

Table 1: Class Characteristics Summary

Class	**Frequency (Hz)**	**Rhythm**	**QRS Pattern**	**Noise Level**
Normal	1.2	Regular	Periodic sharp spikes	Low
Arrhythmia	1.2 ± 0.3	Irregular	Variable intervals	Medium
Fibrillation	2.5–4.0	Chaotic	Absent/degraded	High
Tachycardia	2.0	Regular	Rapid spikes	Medium
Bradycardia	0.8	Regular	Slow spikes	Low

The datasets analysed during the current study are not publicly available due to patient confidentiality and institutional restrictions, but are available from the corresponding author, who is part of the hospital's technical biomedical staff, upon reasonable request.
All methods were carried out in accordance with relevant guidelines and regulations. This study did not involve direct experimental interventions on patients. Instead, the dataset was obtained from anonymized clinical records and monitoring systems within a university hospital in Tunisia and subsequently processed through simulation techniques for research purposes. Therefore, no additional patient intervention or experimentation was performed.

4 Preprocessing Module and Features Extraction

To equalize the signals across all samples, the normalization procedure was used to enhance signal quality and improve model robustness.

After that, the ECG recordings were divided into segments so that the models could examine regular data points.
The process implements two parallel approaches for ECG analysis:
Handcrafted Feature Extraction for ML: RF/SVM
Automatic Feature Learning for DL: CNN/LSTM/BiLSTM

4.1 Features Extraction Analysis: Handcrafted Features (ML)

In order to identify clinically significant traits, the extract features algorithm calculates 15 manufactured features from raw ECG data (Table 2):

Table 2: Handcrafted Feature Categories

Category	**Features**	**Clinical Relevance**
Statistical	Mean, Standard Deviation, Max, Min, Skewness, Kurtosis	Describes signal distribution and shape
Spectral	Dominant Frequency (via FFT)	Identifies abnormal rhythms (e.g., high frequency in fibrillation)
Energy	Root Mean Square (RMS), Total Energy	Measures signal strength and variability
Temporal	25th/75th Percentiles, Interquartile Range (IQR), Peak-to-Peak Amplitude	Quantifies amplitude variations over time
Complexity	Zero-Crossing Rate	Indicates rhythm regularity (e.g., irregular crossings in arrhythmia)

Figure 2 presents the distribution of six different statistical features extracted from the cardiac signal data with different classes.

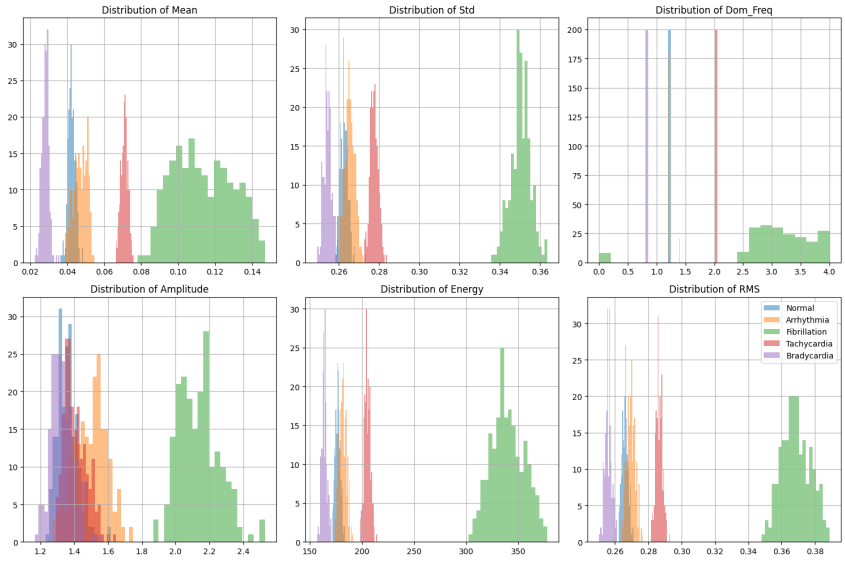

Figure 2: Distribution of ECG features showing how normal signals (blue) clearly differ from cardiac anomalies: bradycardia (purple) exhibits low dominant frequency (0.5 Hz), tachycardia (red) shows high frequency (2 Hz), while fibrillation (green) is characterized by significantly higher values in mean, standard deviation, amplitude, and energy, enabling automated detection of cardiac pathologies

The cardiac diseases are clearly distinguished from one another across a variety of characteristics in the showed histograms. As an example, Fibrillation (green) has much higher Mean, Std, Amplitude, Energy, and RMS values, and the Dominant Frequency distributions display close, condition-specific clusters with little overlap.

4.2 Feature Learning Analysis: Automatic Feature (DL)

Two main architectures are used by deep learning models to interpret raw ECG data. Each of them extracts unique feature representations that are essential for identifying cardiac patterns.
To identify localized morphological traits including QRS complexes and T-wave morphologies, the Convolutional Neural Network (CNN) architecture uses 1D convolutional layers (Conv1D) with learnt kernel filters. While following max-pooling techniques reduce spatial redundancy and maintain diagnostically significant characteristics, these layers detect small waveform changes and spike locations.
CNNs automatically learn to represent important morphological signatures, such as P-wave abnormalities and ST-segment deviations, through hierarchical processing. These signatures are crucial for identifying structural cardiac defects.
In an other hand, Long Short-Term Memory (LSTM) networks, along with Bidirectional LSTM (BiLSTM) models, are specialized in simulating the temporal dynamics of ECG sequences. Heart rate turbulence patterns typical of arrhythmias are captured by these recurrent structures through the analysis of rhythm abnormalities and R-R interval variability. The strong identification of abnormal rhythms, such as atrial fibrillation, is enabled by the bidirectional configuration, which contextualizes each timestep through both forward (past-to-future) and backward (future-to-past) processing.

4.3 Comparative Features Analysis

In order to directly compare conventional machine learning (ML) and deep learning (DL) techniques, the ECG data system precisely matches class specific patterns with both handcrafted and learnt feature representations. (Table 3).

Table 3: Traditional ML vs. Deep Learning Feature Characteristics

Aspect	**Traditional ML**	**Deep Learning**
Features	15 handcrafted features	Raw signal
Feature Source	Manual design	Automatic learning
Strengths	Interpretability	Complex pattern handling
Data Efficiency	Works with small data	Requires large datasets
Clinical Alignment	Direct feature mapping	Learned biomarker discovery

5 ML and DL Training Algorithms

The Random Forest (RF) algorithm, an ensemble learning technique, combines predictions from several randomized decision trees (Figure 3). A baseline sample of the dataset is used to train each tree, and node splits are established by reducing entropy across a random subset of 15 manually created ECG variables (e.g., dominant frequency, mean amplitude). The model maintains interpretability through feature significance metrics while capturing non-linear interactions by introducing randomness through both data and feature sampling. This method works especially well for classifying cardiac rhythms since designed features frequently encode patterns that are clinically significant, including waveform shape or heart rate variability.

In a high-dimensional feature space, the Support Vector Machine (SVM), a maximum-margin classifier, finds the best hyperplane to divide cardiac rhythm classes (Figure 4). The SVM separates complex arrhythmia patterns by projecting the same 15 manually created ECG features that the Random Forest uses into a non-linear space using a radial basis function (RBF) kernel. By using a regularization parameter that accounts for overlapping classes, the model strikes a balance between margin maximization and classification error. Despite being computationally fast, the SVM is less able to adapt to subtle morphological differences in ECG signals because of its dependence on kernel-based modifications, which restricts its capacity to hierarchically learn features from raw data.

Through a hierarchy of 1D convolutional and pooling layers, the Convolutional Neural Network (CNN) architecture automatically extracts localized morphological elements that are essential for rhythm classification from raw ECG signals (Figure 5). Patterns like QRS complexes and P-wave irregularities are gradually captured by three convolutional blocks (64, 128, and 256 filters with kernel size=5), while

max-pooling layers (pool size=2) add translation invariance to slight signal changes. Softmax activation is used in the last dense layers to combine these information into probabilistic class predictions. This architecture is modeled after clinical ECG analysis, where diagnosis is based on spatial patterns rather than temporal sequences.

In order to describe temporal correlations in heart rate and rhythm intervals, the Long Short-Term Memory (LSTM) network uses gated memory cells to analyze ECG signals as sequential data (Figure 6). The sample input is analyzed progressively by two LSTM layers (64 and 128 units), which record associations such abnormal R-R intervals that are suggestive of arrhythmia. The unidirectional architecture is theoretically well-suited for time-series analysis, but because it lacks contextual awareness of future signal segments, it has trouble handling chaotic or aperiodic patterns (like fibrillation). These temporal parameters are mapped to rhythm classes by the last thick layers; however, the model's resilience to noise and signal drift is limited by its sequential processing.

By analyzing ECG signals in both forward and reverse directions, the Bidirectional LSTM (BiLSTM) improves temporal modeling by concurrently incorporating past and future context (Figure 7). By analyzing the input sequence from opposite directions, two bidirectional layers (64 and 128 units) enhance the detection of periodic anomalies such as premature ventricular contractions (PVCs), which rely on the context of the waveforms before and after. Because of its emphasis on temporal coherence, this architecture is still sensitive to signal drift even if it performs better than the conventional LSTM in the classification of bradycardia and fibrillation. Although it comes at a higher processing cost, the BiLSTM's dual-time perspective fills the gap between sequential and spatial analysis.
These figures provide an overview of the technical architecture of all proposed algorithms.

A. Random Forest (RF)

- **Type**: Ensemble learning (bagging)
- **Basics**:
  - Builds 100 decision trees
  - Randomness via bootstrap sampling & feature selection
- **Objective**: Minimize entropy

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2$$

- **Features**: 15 handcrafted (mean, dominant frequency, etc.)
- **Advantages**:
  - Robust to overfitting
  - Handles non-linear relationships

Figure 3: Random Forest Architecture

B. Support Vector Machine (SVM)

- **Type**: Maximum-margin classifier
- **Basics**:
  - Optimal hyperplane: $\mathbf{w}^T\mathbf{x} + b = 0$
  - RBF kernel for non-linear separation
- **Objective**:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

- **Features**: Same 15 as RF
- **Advantages**:
  - Effective in high dimensions
  - Global optimum guaranteed

Figure 4: SVM Architecture

A. Convolutional Neural Network (CNN)

- **Input**: Raw ECG signal
- **1D Convolutional Blocks**:
  - Conv1D(64, kernel=5) + ReLU
  - MaxPooling1D(2)
  - Conv1D(128, kernel=5) + ReLU
  - MaxPooling1D(2)
  - Conv1D(256, kernel=5) + ReLU
  - MaxPooling1D(2)
- **Classifier**:
  - Flatten → Dense(128) + ReLU
  - Dropout(0.5) → Dense(5) + Softmax
- **Rationale**: Hierarchical convolutions detect localized features (QRS complexes); pooling ensures translation invariance.

Figure 5: CNN Architecture

B. Long Short-Term Memory (LSTM)

- **Input**: Raw ECG signal
- **LSTM Blocks**:
  - LSTM(64, return_sequences=True)
  - LSTM(128)
- **Classifier**:
  - Dense(64) + ReLU
  - Dropout(0.3) → Dense(5) + Softmax
- **Rationale**: Sequential processing models temporal dependencies (e.g., R-R intervals) for arrhythmia detection.

Figure 6: LSTM Architecture

C. Bidirectional LSTM (BiLSTM)

- **Input**: Raw ECG signal
- **BiLSTM Blocks**:
  - Bidirectional(LSTM(64, return_sequences=True))
  - Bidirectional(LSTM(128))
- **Classifier**:
  - Dense(64) + ReLU
  - Dropout(0.3) → Dense(5) + Softmax
- **Rationale**: Bidirectional context captures past/future patterns, enhancing detection of periodic anomalies (e.g., premature contractions).

Figure 7: BiLSTM Architecture

6 Perturbation Generation Methodology

The implementation composed five clinically-relevant perturbations through mathematical transformations of the original ECG signals $x(t)$. Each perturbation type is applied to all classes uniformly to test cross-condition robustness.

6.1 Gaussian Noise

Simulates amplifier noise and poor electrode contact. It is presented by the following equation:

$$x_{\text{noisy}}(t) = x(t) + \mathcal{N}(0, \sigma^2) \tag{2}$$

With

- $\sigma \in \{0.05, 0.1, 0.2, 0.3\}$ scaled to signal amplitude
- Additive white noise across all temporal positions
- Simulates poor electrode contact and amplifier artifacts

6.2 Baseline Drift

Caused by respiratory movements and skin impedance changes as follows:

$$x_{\text{drift}}(t) = x(t) + A \sin\left(\frac{2\pi t}{T}\right) \tag{3}$$

Where

- $A \in \{0.1, 0.2, 0.5, 1.0\}$ (normalized amplitude)
- Slow oscillation ($T = 5$ sec) mimics respiratory artifacts
- Phase-aligned across samples for consistent testing

6.3 Missing Data

Represents temporary signal loss during electrode displacement:

$$x_{\text{missing}}(t) = \begin{cases} 0 & \text{with probability } \rho \\ x(t) & \text{otherwise} \end{cases} \tag{4}$$

With

- Missing ratios $\rho \in \{5\%, 10\%, 20\%, 30\%\}$
- Random temporal masking with uniform distribution
- Simulates transient signal loss from motion artifacts

6.4 Heart Rate Variation

Tests robustness to pathological bradycardia/tachycardia presented as follows:

$$x_{\text{scaled}}(t) = \text{resample}(x(\alpha t)) \tag{5}$$

Where

- Scaling factors $\alpha \in \{0.8, 0.9, 1.1, 1.2\}$
- Linear interpolation for time-stretch/compression
- Maintains original signal length through truncation/padding

6.5 Motion Artifacts

Mimics muscle activity and sudden movement interference:

$$x_{\text{motion}}(t) = x(t) + \underbrace{\sum_{i=1}^{N_s} A_s \delta(t - t_i)}_{\text{spikes}} + \underbrace{A_m \sin(2\pi f_m t)}_{\text{high-freq}} \tag{6}$$

With

- Spike parameters: $A_s = 2\sigma_x$, $N_s = 0.02N_{\text{samples}}$
- High-frequency: $A_m \in \{0.1, 0.2, 0.4, 0.6\}$, $f_m = 25$ Hz
- Combines transient spikes and muscle artifact simulation

All perturbations applied to normalized signals ($\mu = 0$, $\sigma = 1$) seed=42 for reproducible randomization. Perturbation parameters chosen based on clinical literature. Each perturbation tested independently (no combined effects).

The dataset used includes ECG recordings categorized into five distinct classes: Normal, Arrhythmia, Fibrillation, Tachycardia and Bradycardia.
The classification results, including accuracy metrics and a thorough performance analysis, are presented and discussed in the next section.
Since the SVM model showed poor accuracy and the Random Forest model runs stochastically, only deep learning models CNN, LSTM, and BiLSTM were assessed for robustness.
Models with inconsistent or subpar performance were deemed unsuitable for this additional testing of robustness due to the critical sensitivity needed in medical cardiac applications. Therefore, we assess how resilient the suggested models are to different obstacles like noise, signal drift, missing data, heart rate fluctuations, and motion artifacts.

7 Results and Discussions

The models were evaluated using standard classification metrics. If:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

1. **The Accuracy** is presented by the following equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2. **The Recall (Sensitivity)** having the equation below:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3. **F1 Score** follows this equation:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

4. **The Precision** which is :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

7.1 Computational Performance

Critical trade-offs for implementing models in real-time ECG analysis are highlighted by the computational efficiency data in Table 4.

Table 4: Computational Efficiency of Models

Model	**Execution Time (min)**	**Inference Time (s)**
RF	3	0.002
SVM	5	0.0318
CNN	20	2.3109
LSTM	60	25.1243
Bi-LSTM	115	112.5476

With quick training (3–5 minutes) and near-instant inference (0.002–0.0318 seconds), traditional models like Random Forest (RF) and SVM are useful for continuous cardiac monitoring, where quick anomaly identification (such as arrhythmia) is effective. But because of their simplicity, they might not be able to pick up on little temporal patterns in ECG waveforms, including irregular heartbeats or variations in the ST-segment.

CNNs, on the other hand, provide a compromise by using spatial feature extraction to detect localized abnormalities while still being practical for near-real-time applications (20-minute training, 2.31 s inference). Although their slower inference restricts deployment in low-latency systems, LSTMs (60-minute training, 25.12s inference) are excellent at modeling temporal dependencies, which makes them appropriate for identifying sequential abnormalities like atrial fibrillation or ventricular tachycardia. The Bi-LSTM may be able to capture bidirectional context for complicated diagnoses, but it runs the danger of being impractical in clinical settings with limited resources due to its lengthy training period (112.5 minutes) and missing inference data.

7.2 Classification Performance

In this comparative analysis of cardiac rhythm classification models, the CNN was the most reliable model in this comparison of cardiac rhythm classification models, with a remarkable accuracy and F1-score with only one instance of an arrhythmia patient being incorrectly classified as normal. This demonstrates how well convolutional networks capture discriminative spatiotemporal patterns from cardiac data (Confusion matrix is illustrated in Figure 8).

Significantly, the Random Forest (Figure 9) model performed similarly, defying expectations by matching the performance of advanced deep learning architectures such as BiLSTM (Figure 10) and the LSTM (Figure 11). This suggests that feature engineering in traditional machine learning is still competitive for physiological data that is well structured.

Added the results of SVM algorithm presented in Figure 12, all models achieved perfect classification for Tachycardia, indicative of its distinct electrophysiological signature, while performance variations in other classes likely reflect differences in arrhythmia complexity and feature separability.

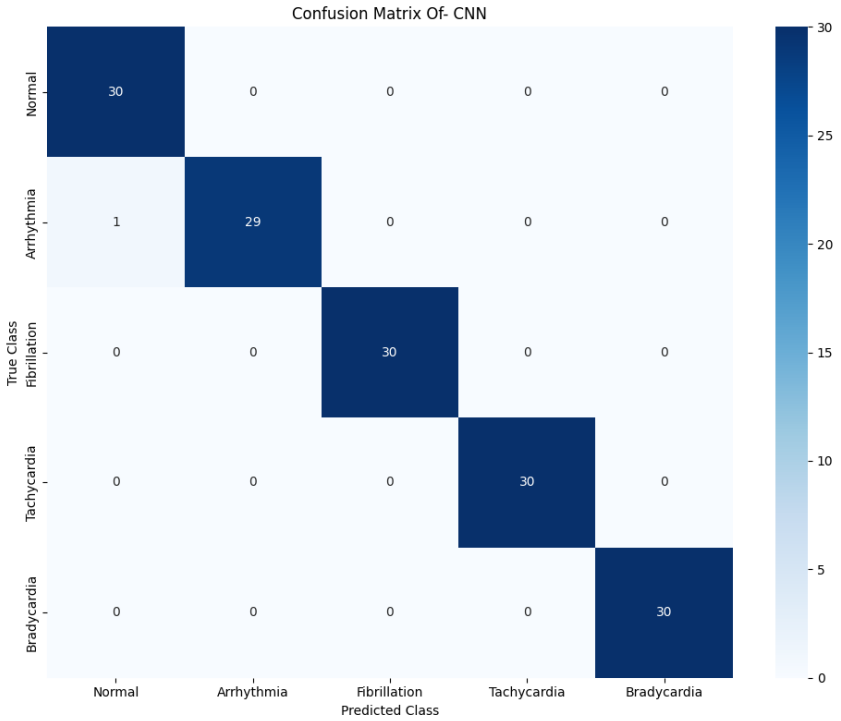

Figure 8: CNN Confusion Matrix

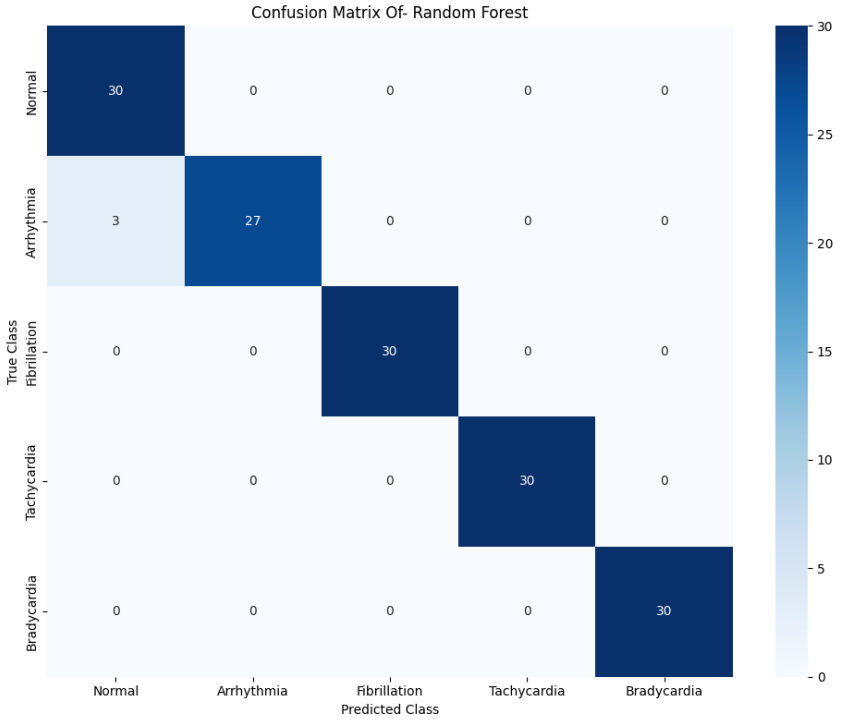

Figure 9: RF Confusion Matrix

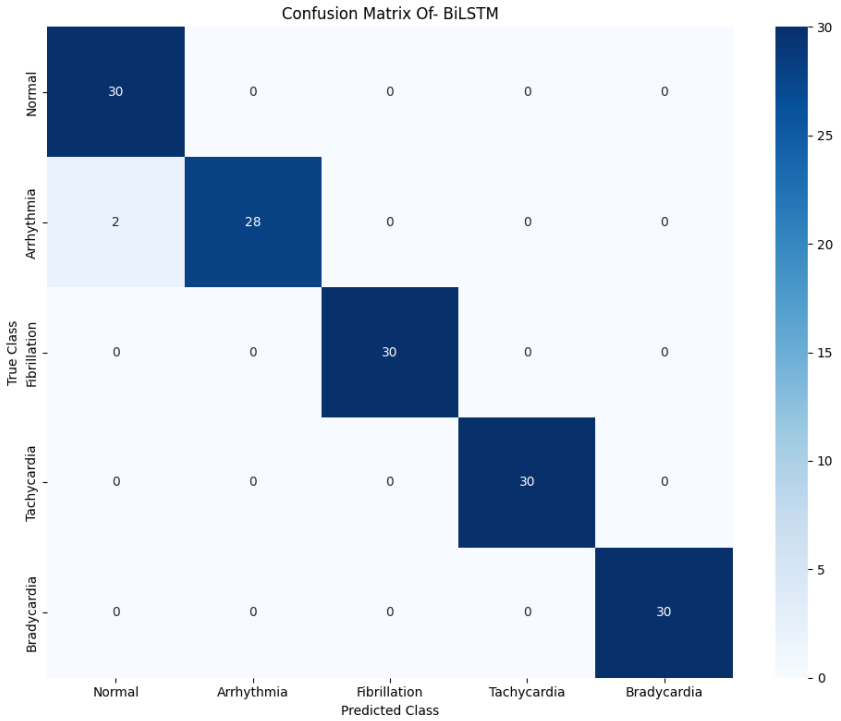

Figure 10: Bi LSTM Confusion Matrix

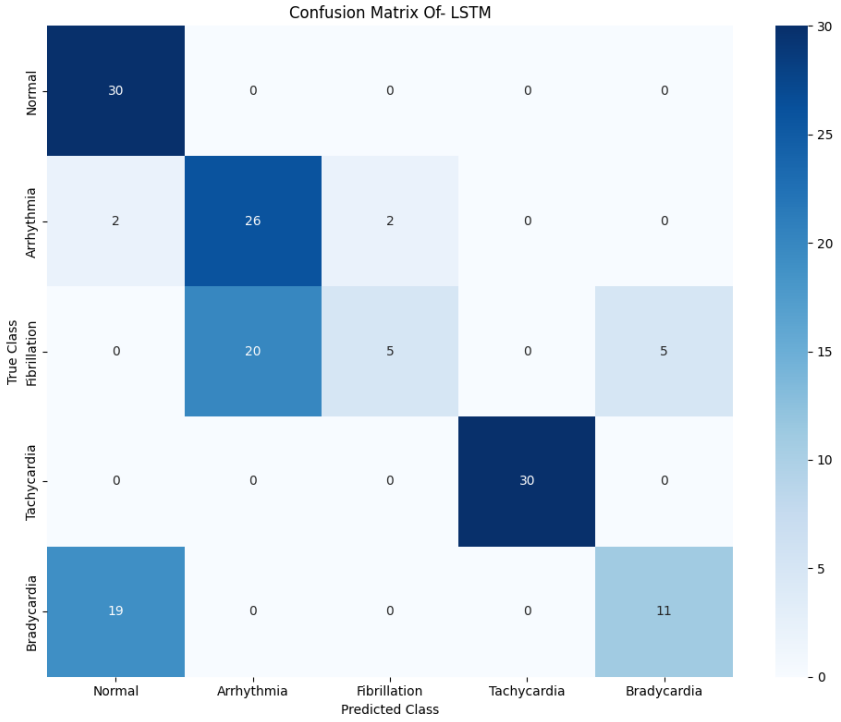

Figure 11: LSTM Confusion Matrix

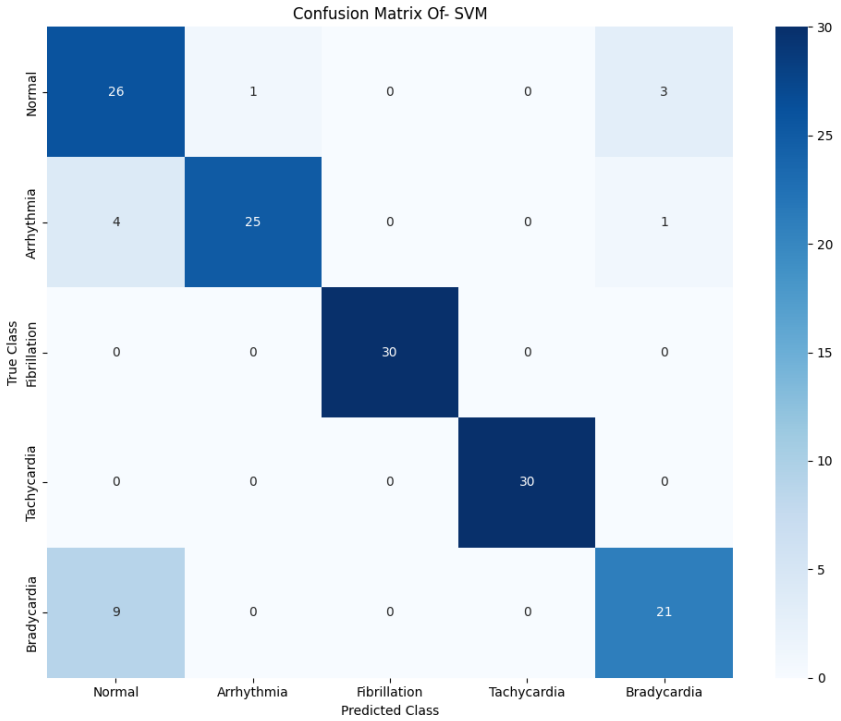

Figure 12: SVM Confusion Matrix

With the greatest results on all parameters (precision: 0.9935, recall: 0.9933, F1-score: 0.9933, accuracy: 0.99), the CNN demonstrates its resilience in maintaining a balance between precision and recall, which is essential for accurate arrhythmia identification.
The Random Forest follows (accuracy: 0.98, F1-score: 0.9799), demonstrating its competitiveness even though it is a conventional machine learning technique. Although the BiLSTM has a great F1-score (0.9867) and nearly perfect accuracy (0.99), it handles ambiguous instances with some inconsistencies, as seen by its slightly poorer precision (0.9875) and recall (0.9867) when compared to CNN.
The SVM, on the other hand, performs moderately (F1-score: 0.8820, accuracy: 0.89), most likely as a result of difficulties with non-linear feature correlations or class overlaps.
The LSTM's weak recall (0.68), which puts it at danger of missing important arrhythmia diagnoses, and its far behind F1-score (0.6347) and accuracy (0.68) demonstrate its insufficiency for this task.
Based on Table 5, the poor recall of the LSTM reveals its unreliability in detecting actual arrhythmia patients, despite the fact that all models save the LSTM attain near perfect accuracy (>0.89), limiting false positives. While Random Forest's impressive performance indicates that feature-engineered techniques are still viable options in clinical scenarios where model interpretability is a top priority, our results highlight CNN architectures' dominance for ECG signal processing. The typical LSTM's looking underperformance highlights the need for architectural improvements for sequential medical data.

Table 5: Overall Performance Metrics

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
CNN	0.9935	0.9933	0.9933	0.99
Random Forest	0.9818	0.9800	0.9799	0.98
BiLSTM	0.9875	0.9867	0.9867	˜0.99
SVM	0.8936	0.8800	0.8820	0.89
LSTM	0.7110	0.6800	0.6347	0.68

Based on Table 6, critical performance differences between models and cardiac rhythm categories are shown by the class-specific F1-scores. With the exception of the LSTM, all models achieve perfect classification (F1-score: 1.00) for Bradycardia, Tachycardia, and Fibrillation. This is in keeping the unique electrophysiological characteristics of each condition, such as the chaotic waveform patterns of Fibrillation and the unmistakable high rate threshold of Tachycardia. However, there is a significant

difference in performance between the Normal and Arrhythmia groups. While the SVM and LSTM models have trouble, especially with normal rhythms (SVM: 0.75, LSTM: 0.74) and arrhythmia (LSTM: 0.68), the CNN, Random Forest, and BiLSTM models show good consistency (F1-scores: 0.95–0.98).
The LSTM's poor performance in Bradycardia (0.48) indicates insufficient sensitivity to low rate patterns, while its disastrous failure in Fibrillation classification (F1-score: 0.27) highlights its incapacity to model irregular rhythm dynamics without bidirectional context.

Table 6: Class-specific F1-Scores

Model	**Normal**	**Arrhythmia**	**Fibrillation**	**Tachycardia**	**Bradycardia**
CNN	0.98	0.98	1.00	1.00	1.00
Random Forest	0.95	0.95	1.00	1.00	1.00
BiLSTM	0.97	0.97	1.00	1.00	1.00
SVM	0.75	0.89	1.00	1.00	0.76
LSTM	0.74	0.68	0.27	1.00	0.48

7.3 ROC Curves

For most of the assessed models, the ROC curves show remarkable discriminative capability.
For all five cardiac situations, the BiLSTM (Figure 13) and CNN (Figure 14) designs show ideal area under the curve (AUC = 1.00), with their curves closely hugging the plots' upper-left corner. This suggests that at almost any classification threshold, the model may attain perfect sensitivity without compromising specificity.
Similar to this, Random Forest (Figure 15) performs exceptionally well, achieving near perfect AUC values between 0.99 and 1.00 in all situations. Regardless of the decision threshold selected, its ROC curves indicate strong discriminative skills with little departure from the ideal. Despite having more noticeable step-like progressions in its curves, the SVM model (Figure 16) also performs well, obtaining perfect AUC values (1.00) for the classifications of Bradycardia (0.98), Arrhythmia (0.96), and Normal (0.94), as well as ideal scores for Fibrillation and Tachycardia.
The ROC characteristics of the LSTM model (Figure 17) exhibit more inconsistent performance. It performs somewhat worse for Bradycardia (0.96), Arrhythmia (0.93), and Fibrillation (0.92), but achieves flawless discrimination for Normal and Tachycardia situations (AUC = 1.00).
Greater concavity in the curves under these situations suggests more noticeable trade-offs between sensitivity and specificity across various thresholds.

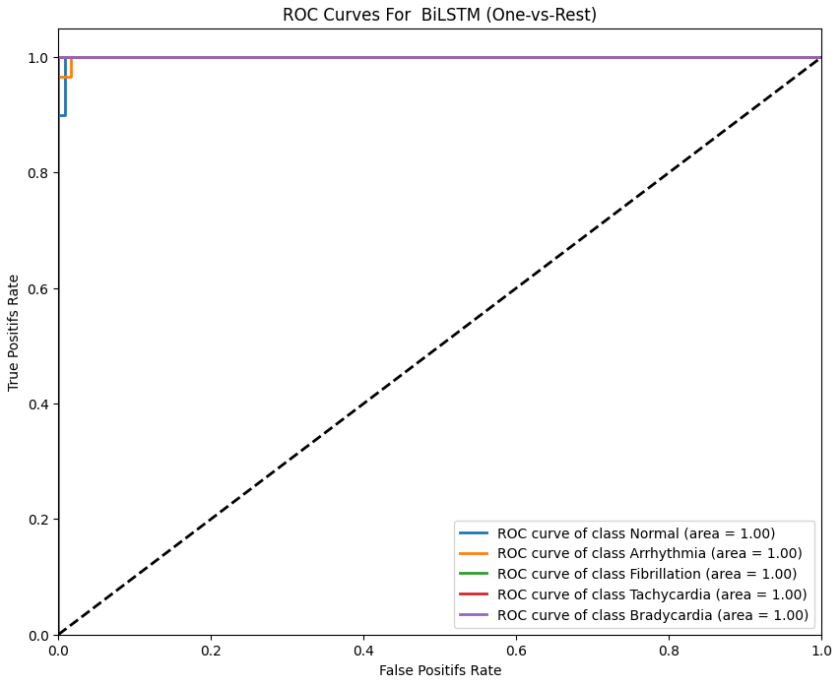

Figure 13: ROC curves for BiLSTM model showing perfect classification performance (AUC=1.00) for all five cardiac conditions in a one-vs-rest framework, demonstrating the exceptional capability of bidirectional LSTM in distinguishing between normal and pathological ECG patterns.

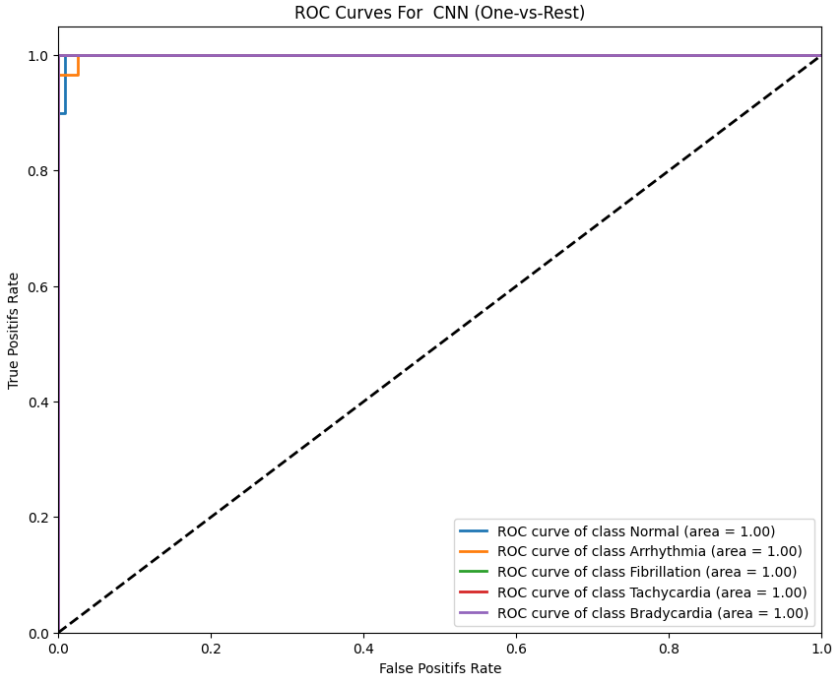

Figure 14: ROC curves for CNN model revealing perfect discrimination (AUC=1.00) across all cardiac conditions, confirming convolutional neural networks' effectiveness in capturing spatial features within ECG signals for accurate cardiac anomaly detection.

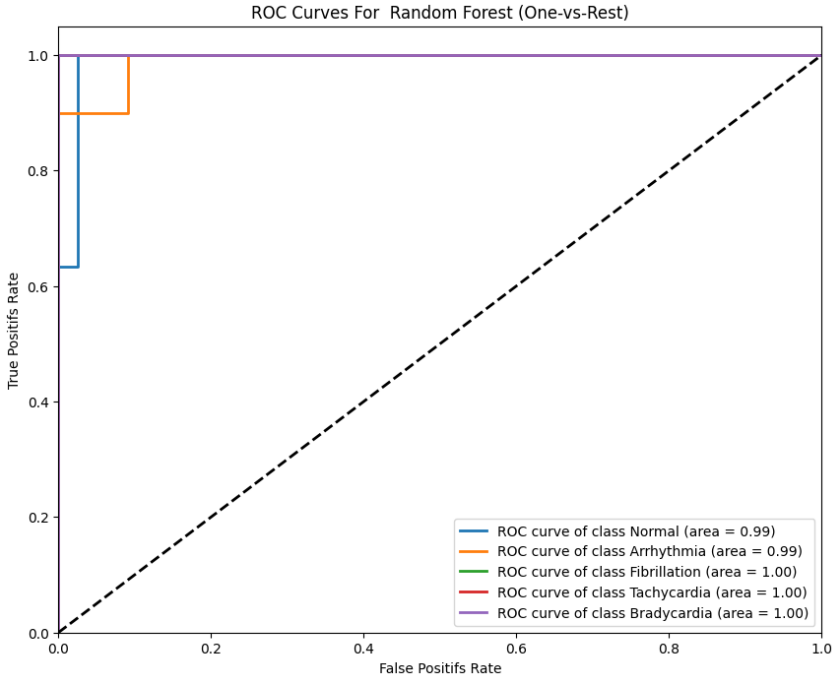

Figure 15: ROC curves for Random Forest classifier showing near-perfect performance (AUC0.99) for all cardiac conditions, demonstrating this ensemble method's robust ability to separate normal from pathological ECG patterns using decision trees

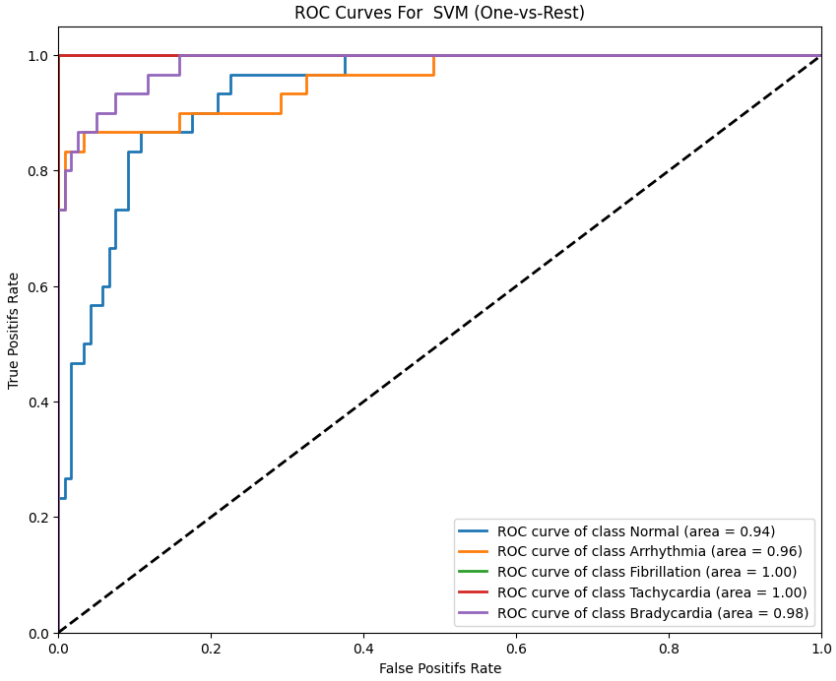

Figure 16: ROC curves for SVM classifier displaying strong but slightly varied performance across cardiac conditions (AUC from 0.94 to 1.00), with perfect classification for Fibrillation and Tachycardia but slightly lower accuracy for Normal class.

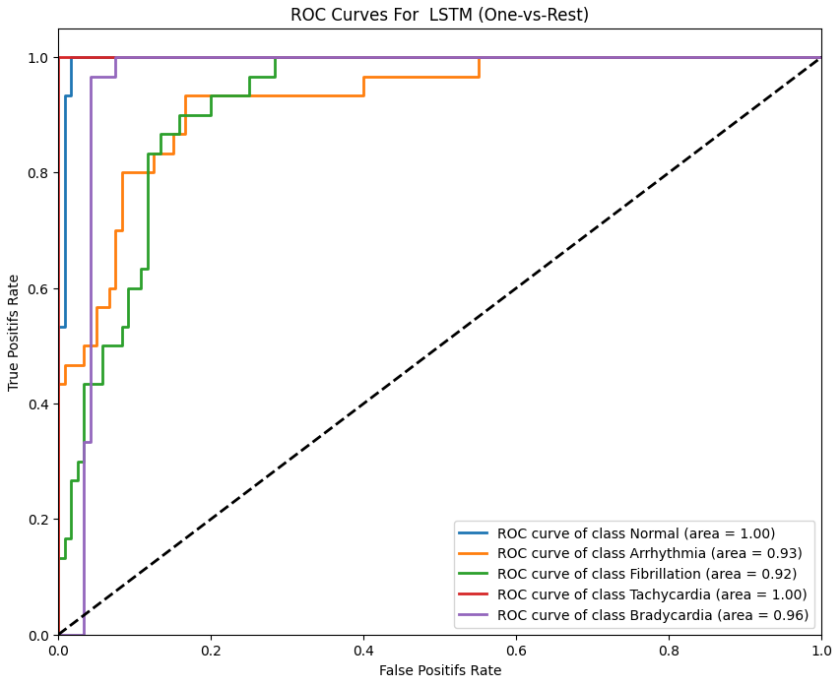

Figure 17: ROC curves for LSTM model showing excellent but varied performance across cardiac conditions, with perfect classification for Normal and Tachycardia (AUC=1.00) but slightly reduced accuracy for Arrhythmia (AUC=0.93) and Fibrillation (AUC=0.92).

The precision-recall curves provide complementary insights into model performance, particularly for the imbalanced classification scenario often encountered in medical diagnostics.

Once again, BiLSTM (Figure 18) and CNN (Figure 19) show outstanding performance, achieving Average Precision (AP) values of 1.00 for all cardiac diseases. Even despite collecting almost all positive occurrences, their precision-recall curves show uncommon erroneous positive predictions, maintaining maximum precision until very high recall levels.

With near perfect scores for Normal (0.96) and Arrhythmia (0.97) conditions and perfect AP scores (1.00) for Fibrillation, Tachycardia, and Bradycardia, Random Forest (Figure 20) exhibits exceptional precision recall characteristics. At the highest recall thresholds, its curves show only slight precision degradation.

More variety is shown by the SVM model (Figure 21), which performs flawlessly for tachycardia and fibrillation (AP = 1.00), performs well for bradycardia (0.95) and arrhythmia (0.92), but maintains precision for normal rhythm (0.79) noticeably worse, especially at larger recall levels.

The greatest variable precision recall performance is shown by LSTM (Figure 22). It performs well with normal rhythm (0.98) and excels at classifying tachycardia (AP = 1.00), but it has significant trouble sustaining accuracy for arrhythmia (0.80), fibrillation (0.70), and bradycardia (0.70). The significant decrease in accuracy as recall values rise for these circumstances points to difficulties differentiating their unique temporal patterns from those of other cardiac states.

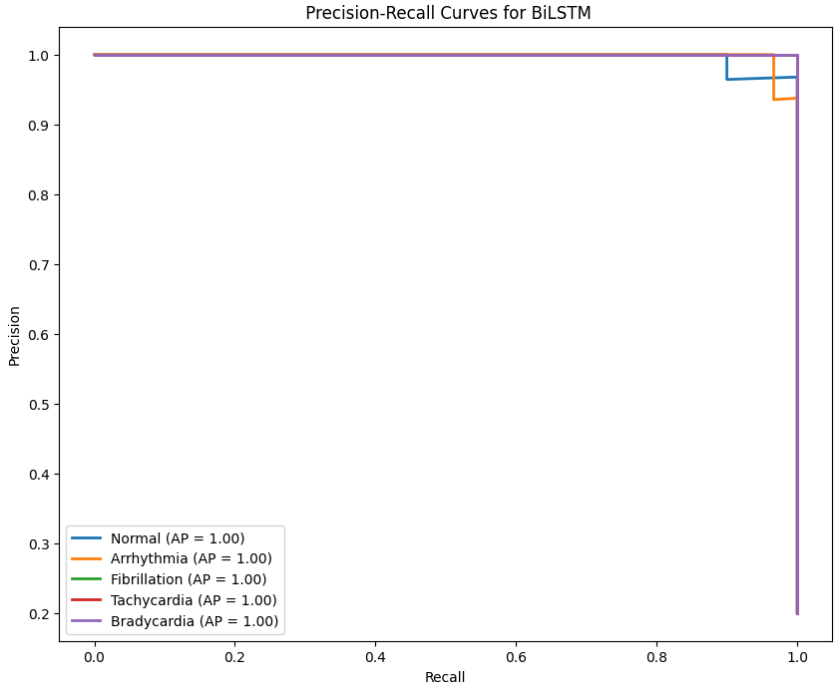

Figure 18: Precision-Recall curves for BiLSTM model showing perfect classification performance (AP=1.00) for all cardiac conditions, with precision maintained at 1.0 across nearly all recall thresholds, only dropping slightly at the highest recall levels, demonstrating the model's exceptional ability to correctly identify cardiac anomalies with minimal false positives.

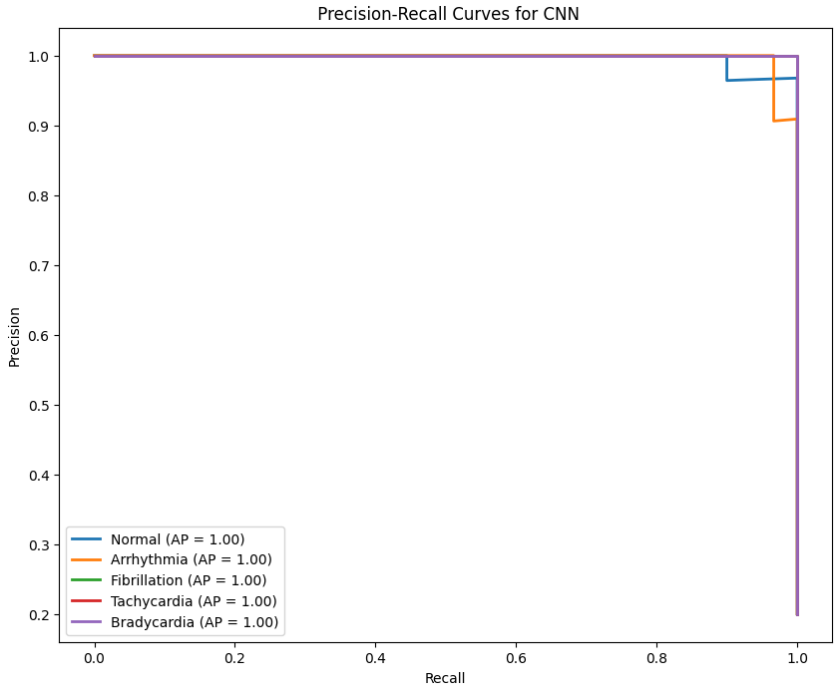

Figure 19: Precision-Recall curves for CNN model exhibiting perfect average precision (AP=1.00) for all cardiac conditions, maintaining maximum precision until very high recall values, confirming CNN's powerful feature extraction capabilities for ECG signal classification with negligible false positive predictions.

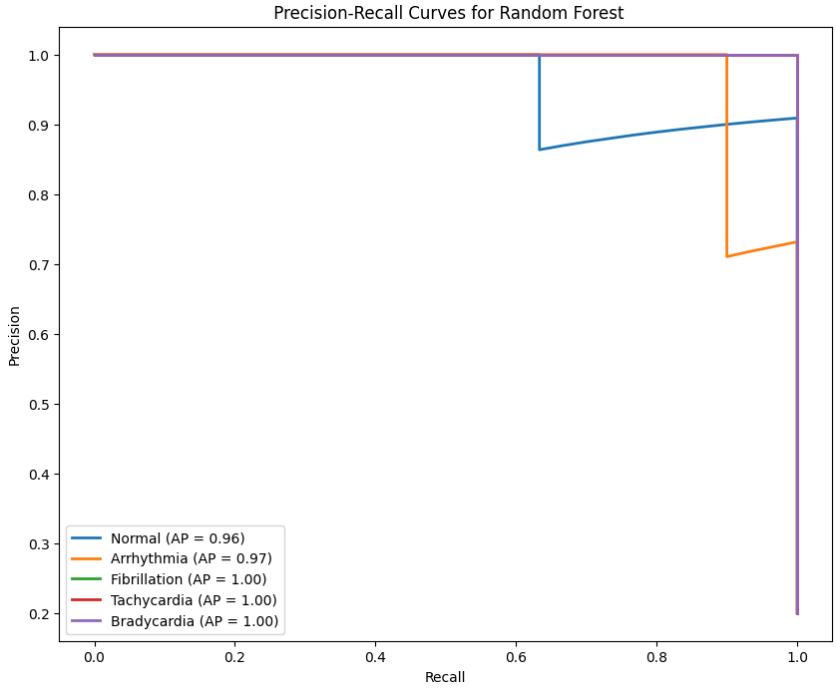

Figure 20: Precision-Recall curves for Random Forest classifier showing perfect performance for Fibrillation, Tachycardia and Bradycardia (AP=1.00), with slightly lower but still excellent performance for Normal (AP=0.96) and Arrhythmia (AP=0.97) classes, demonstrating the ensemble method's strong classification abilities.

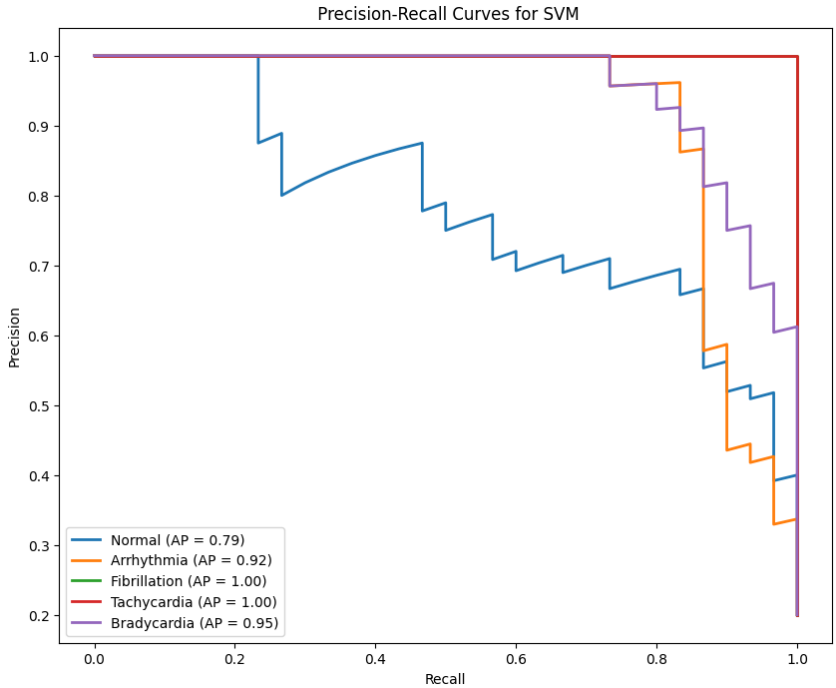

Figure 21: Precision-Recall curves for SVM model revealing varied performance across cardiac conditions, with perfect prediction for Fibrillation and Tachycardia (AP=1.00), strong performance for Bradycardia (AP=0.95) and Arrhythmia (AP=0.92), but notably lower accuracy for Normal class (AP=0.79), illustrating classification challenges for normal heart rhythms.

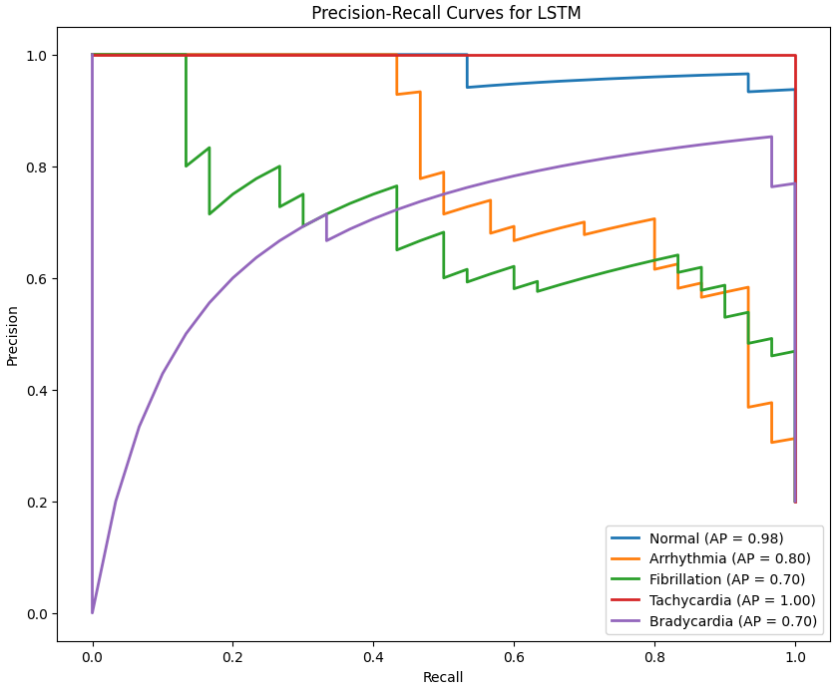

Figure 22: Precision-Recall curves for LSTM model showing excellent performance for Tachycardia (AP=1.00) and Normal class (AP=0.98), but significantly lower performance for Arrhythmia (AP=0.80), Fibrillation and Bradycardia (both AP=0.70), revealing the model's inconsistent classification ability across different cardiac conditions.

The results show that architectural complexity does not necessarily correlate with performance for this specific cardiac classification task, as evidenced by the strong showing of the Random Forest model. Overall, the evaluation metrics show that BiLSTM and CNN provide the most reliable classification across all cardiac conditions, with Random Forest coming in close behind. SVM offers strong performance with some condition-specific limitations, while LSTM shows more inconsistent effectiveness across the various cardiac conditions.
The next section focuses on evaluating the robustness of the deep learning models.

7.4 Robustness Study

To begin with the Figures (23, 24 and 25) that present the training dynamics for three neural network architectures (BiLSTM, CNN, and LSTM) repectively, tracking both accuracy and loss metrics over approximately 19 epochs.
With a high validation accuracy (98%) that holds steady during training and a rapid achievement and maintenance of low loss values after just two to three epochs, the CNN model exhibits exceptional stability and performance.
The BiLSTM, on the other hand, exhibits decent but more erratic learning, settling at close to 100% accuracy by the last epochs after a dramatic accuracy spike at epoch 7.5 and a similar validation loss spike.
The LSTM model most prominently displays catastrophic forgetting about epoch 15, when training and validation accuracy sharply decline from approximately 98% to less than 40%, accompanied by comparable increases in loss, before exhibiting a little rebound in the last epochs.

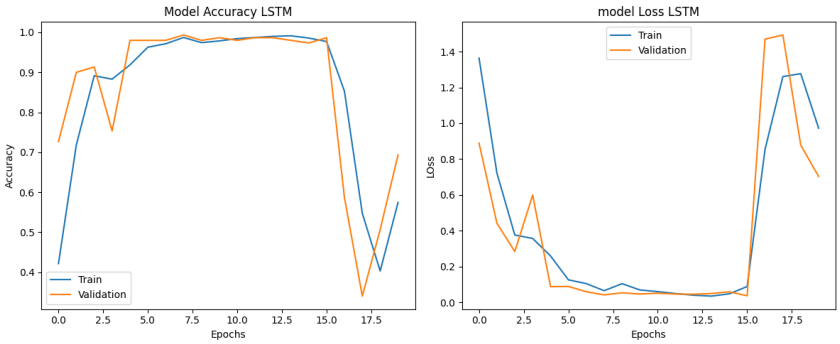

Figure 23: LSTM Training Curves

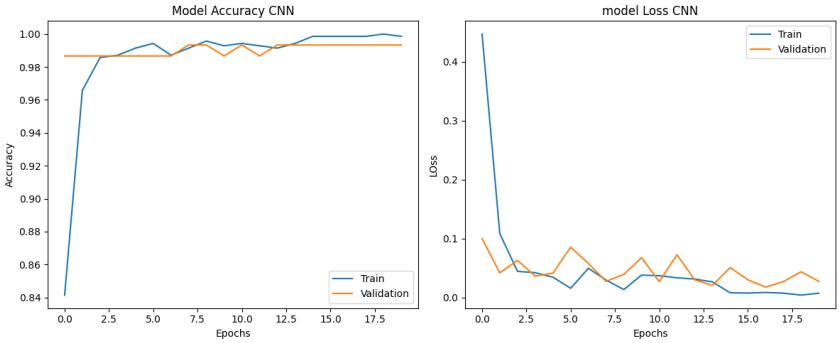

Figure 24: CNN Training Curves

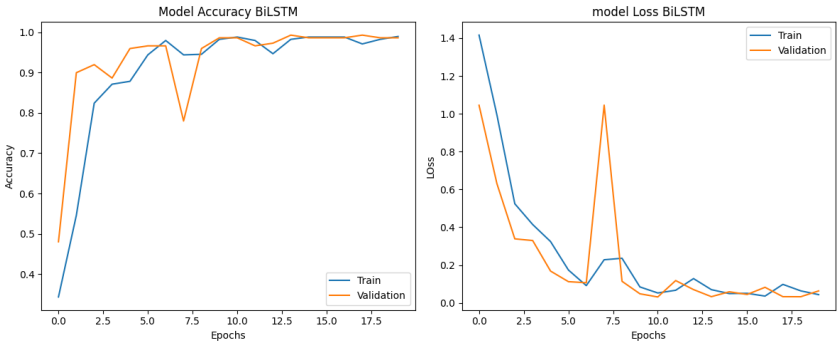

Figure 25: Bi LSTM Training Curves

Model Robustness visualizations are explained next. This robustness evaluation, incorporating sensitivity to noise, missing data, artifacts, heart rate variations, and signal drift, reveals critical insights into the practical viability of the tested models for real-world ECG classification.

Robustness against Missing Data Perturbation

Figure 26 illustrates the effect of missing data perturbation on signal processing and model performance. The top graph displays the original signal data subjected to increasing percentages of missing data (5%, 10%, 20%, and 30%).

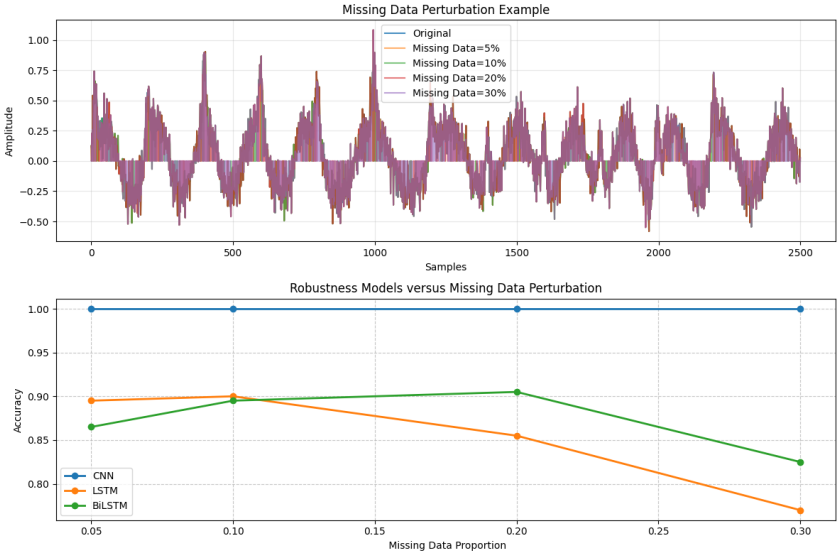

Figure 26: Missing Data Evaluation Models

Where data points are absent, obvious gaps show up, but the general waveform pattern is still identifiable. The response of models to this missing data is seen in the bottom graph. The CNN model has exceptional resilience, preserving perfect accuracy (1.0) at all missing data levels. The BiLSTM model, on the other hand, first reaches around 86% accuracy at 5% missing data, then gradually improves between 10% and 20%, before gradually declining to about 82% accuracy at 30% missing data. With an initial accuracy of almost 89%, the LSTM model's performance gradually declines as the fraction of missing data rises. It falls more precipitously than BiLSTM and reaches roughly 77% at 30% missing data.

Robustness against Gaussian Noise Perturbation

The effect of noise perturbation on the signal and model performance is shown in Figure 27.

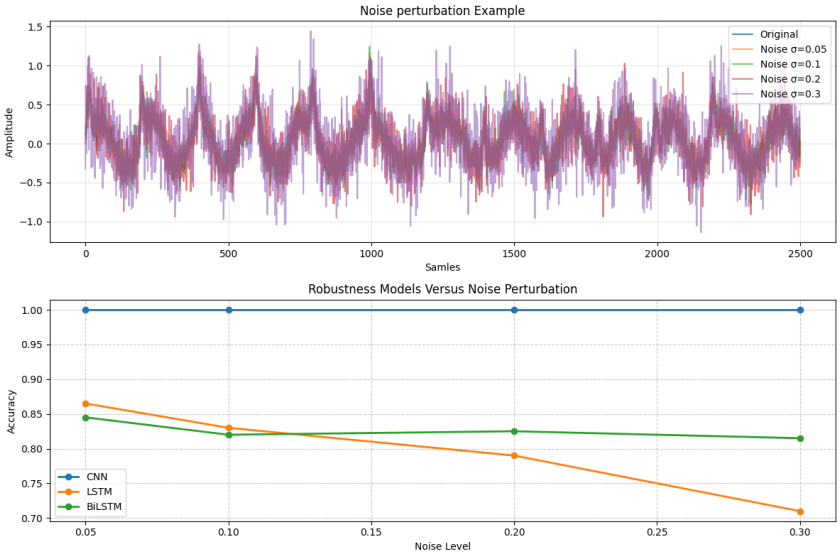

Figure 27: Noise Evaluation Models

As the noise levels (σ = 0.05, 0.1, 0.2, and 0.3) increase, the original signal is displayed in the top graph, where the waveform is increasingly distorted and jagged. The models' accuracy responses are shown in the bottom graph: the CNN maintains perfect accuracy (1.0) at all noise levels, but the LSTM begins at around 86% accuracy at low noise levels and gradually decreases to about 70% at the maximum noise level. In spite of the growing noise, the BiLSTM exhibits better stability, keeping accuracy between 81% and 84%.

Robustness against Motion Artifact Perturbation

The impact of motion artifact perturbation on the signal and model performance is seen in Figure 28.

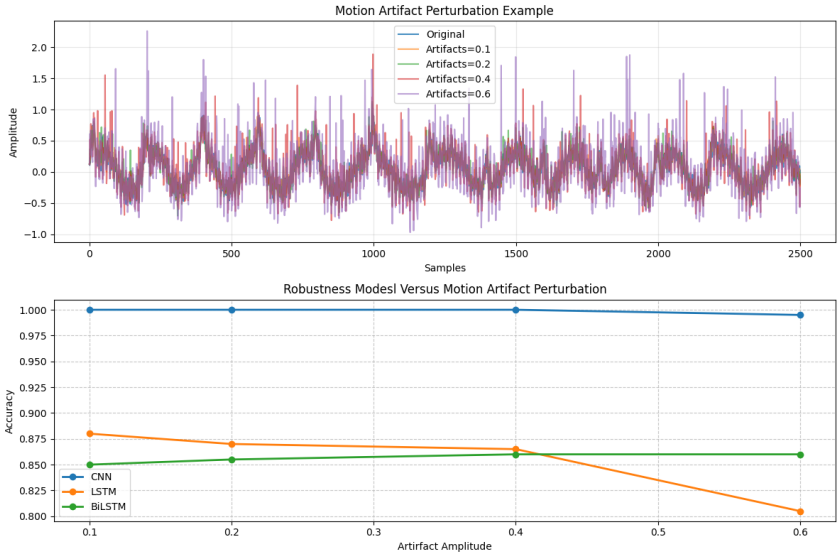

Figure 28: Motion Artifact Evaluation Models

The top graph displays the original signal with abrupt spikes and abnormalities caused by increasing motion artifact amplitudes (0.1, 0.2, 0.4, and 0.6). The model responses are shown in the bottom graph: the CNN maintains near-perfect accuracy (99%) even at the highest artifact level, the LSTM begins at about 88% accuracy and progressively drops to about 80% at an artifact amplitude of 0.6, and the BiLSTM shows remarkable stability, maintaining or even marginally improving its accuracy across various artifact levels.

Robustness against Drift Baseline Perturbation

Figure 29 highlights the impact of baseline drift perturbation on the signal and model performance.

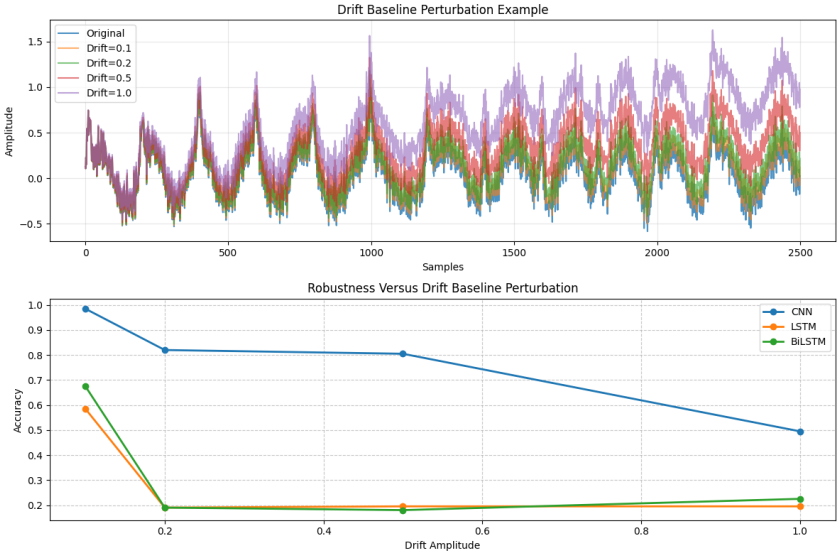

Figure 29: Drift Evaluation Models

Both models identify baseline drift as the most destructive type of perturbation. The top graph displays the original signal with increasing drift levels (0.1, 0.2, 0.5, 1.0), resulting in a progressive vertical displacement over time. The bottom graph shows that drift causes the most dramatic decline in model accuracy: the CNN, initially at around 98% accuracy, drops sharply to about 50% at the highest drift level, while both LSTM and BiLSTM experience catastrophic degradation starting from drift levels

of 0.2, falling to about 20% accuracy.

Robustness against Heart Rate Perturbation

The impact of heart rate perturbation on the signal and model performance is shown in Figure 30.

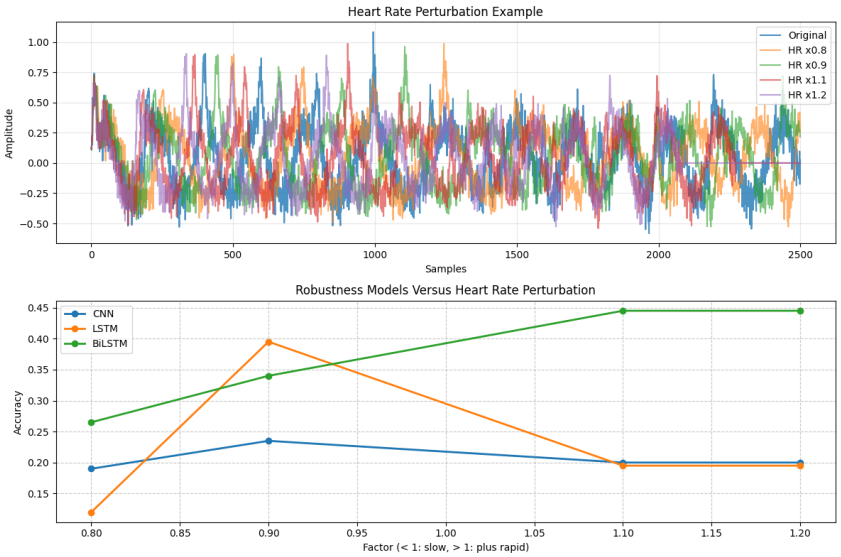

Figure 30: Heart Rate Evaluation Models

The top graph displays the original signal after its frequency was changed by heart rate fluctuations (x0.8, x0.9, x1.1, x1.2). Under these circumstances, the bottom graph shows typically poor model performance: regardless of heart rate variations, all models show relatively low accuracy (20–45%). While LSTM exhibits erratic performance with a peak at x0.9, BiLSTM outperforms them all, especially at higher heart rates (x1.1 and x1.2). At about 20% accuracy, the CNN is still comparatively steady but continuously low.

Global Robustness Evaluation Models

Figures 31, 32 and 33 conclude all the evaluation robustness metrics of the proposed CNN, LSTM and Bi LSTM models.

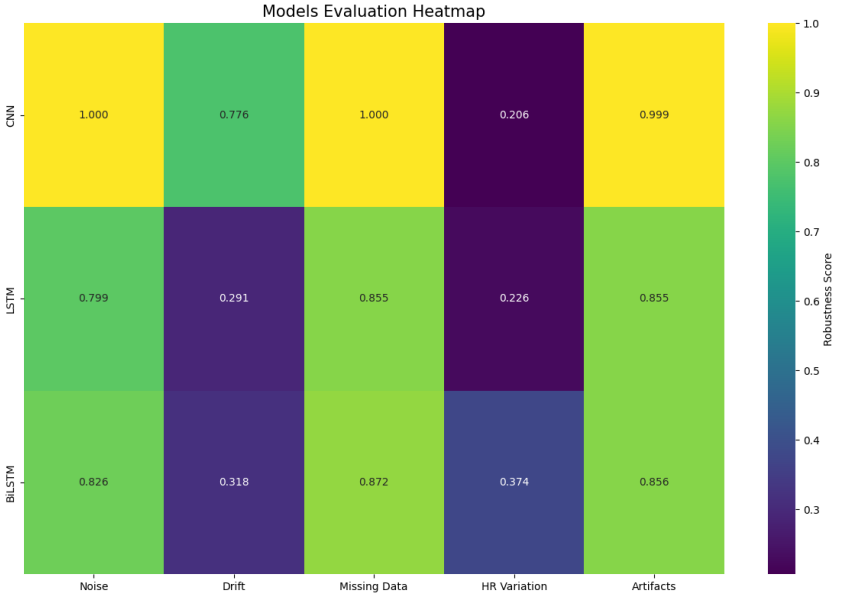

Figure 31: Models Heatmap

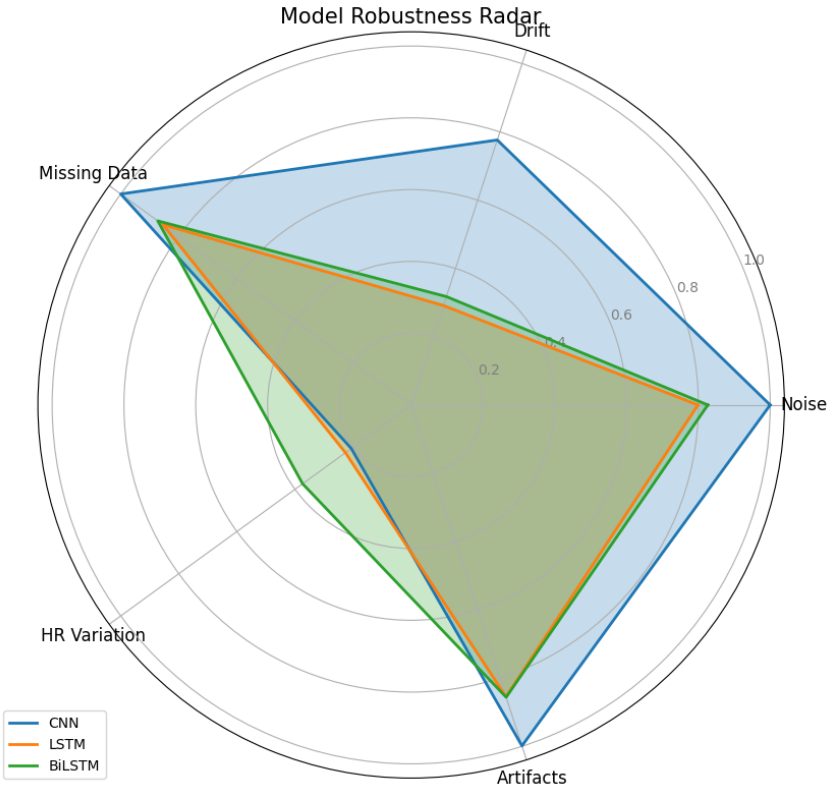

Figure 32: Models Radar

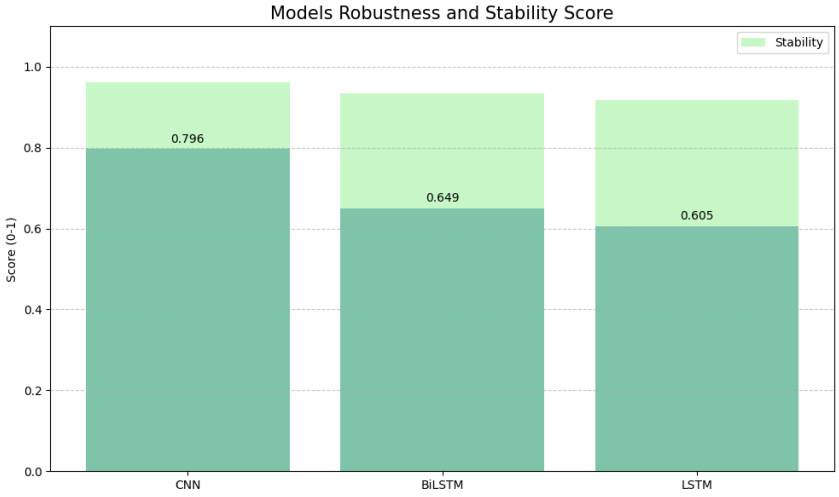

Figure 33: Models Stability

The CNN's extreme susceptibility to heart rate fluctuations (0.206) highlights a crucial limitation, even though it obtained the highest overall robustness score (0.796). This is probably because it relies on fixed convolutional kernels that are optimized for static waveform morphologies, which causes even small heart rate fluctuations to impair its performance.

The LSTM's low resilience (0.605), especially its vulnerability to signal drift (0.291), on the other hand, is consistent with its previous classification failures and highlights its incapacity to represent long-term temporal relationships in clinical data that is noisy.

The BiLSTM achieves a moderate level of resilience (0.649), with balanced tolerance to artifacts (0.856) and noise (0.826). However, its difficulties with drift (0.318) and heart rate fluctuations (0.374) reflect constraints observed in all designs.

These results contribute to the explanation of the previous model classification performances. Although the CNN is quite accurate when dealing with clean data, it clearly falters when dealing with heart rate variability, which is a frequent problem in actual ECG monitoring. The intermediate robustness of the BiLSTM is consistent with its usually resilient but somewhat erratic behavior, indicating that bidirectional processing aids in reducing noise mistakes but does not completely eliminate them. The LSTM seems inappropriate for clinical usage due to its low classification performance and susceptibility to drift.

The inability of all models to handle changes in heart rate indicates a basic flaw in the architectures used nowadays. Heart rate dynamics are not explicitly modeled by either LSTMs or convolutional layers, which place more emphasis on sequential dependencies and specific waveform characteristics (such as ST segments). Future research may incorporate heart rate-adaptive systems like attention modules or dynamic time-warping layers.

The overall subpar performance against signal drift (scores: 0.291–0.776) is consistent with clinical situations in which recordings are tainted by patient or electrode movements. This emphasizes the necessity of using adversarial training or preprocessing pipelines with drift-removal methods to mimic drift during model training.

The variability of ECG signals can impact signal features because they are impacted by a number of parameters, including age, sex, and physiological or psychological states. These factors are not included in this study, which might reduce the precision of the suggested categorization method.

Future research might use the Tempered Fractional Gradient Descent (TFGD) method, which was introduced by Naifar et al. [21], to further improve the optimization framework and speed up convergence in clinical anomalies prediction. In particular, for noisy, high-dimensional ECG data, the model could achieve faster stabilization and higher accuracy by substituting TFGD's tempered memory mechanism, which balances historical gradient contributions via fractional coefficients (α) and exponential decay (λ), for classical gradient descent in the MLP architecture.
This method is in line with the requirement for reliable optimization in dynamic clinical settings and has been proven in medical classification tasks.

8 Conclusion

According to our research, CNN is the best model for classification tasks, exhibiting remarkable accuracy and resilience to data irregularities. RF, SVM, LSTM and BiLSTM models performed competitively, although their precision and recall were not as good as CNN's.
According to the robustness research, CNN constantly outperforms the other models (LSTM and Bi LSTM) in terms of performance against various kinds of data inconsistencies. The stability scores emphasize CNN's appropriateness for real-world applications and further justify its dependability over competing systems.
Our findings highlight how crucial it is to select models according to certain standards like accuracy, stability, and robustness when creating efficient machine learning solutions for categorization issues.
Potential improvements to LSTM and BiLSTM designs to boost their dependability and performance should be investigated in future studies.

9 Declaration

Funding : This research received no external funding.
Competing Interests : The authors declare that they have no competing interests.
Ethics Approval :The research protocol was reviewed and approved by the Ethics Committee of the Middle WestRegional Hospital, Tunisia, which operates under the authority of the Ministry of Health, Tunisia. The data analyzed in this study consisted of anonymized clinical information obtained from the emergency Staff. The first author is a member of the biomedical staff at the Middle West Regional Hospital, and the data were obtained from the historical machine with .D25 files, without any experimental

intervention or modification of patient care. Accordingly, the Ethics Committee waived the requirement for individual informed consent, as the study involved only secondary analysis of de-identified clinical data in accordance with national ethical regulations. All patient data were handled confidentially and used solely for research purposes.
Consent to Participate : Not applicable
Consent to Publish : Not applicable
Data Availability Statement : The datasets analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Gaziano, Thomas A. "Cardiovascular diseases worldwide." Public Health Approach Cardiovasc. Dis. Prev. Manag 1 (2022): 8-18.

[2] Xie, Liping, et al. "Computational diagnostic techniques for electrocardiogram signal analysis." Sensors 20.21 (2020): 6318.

[3] Hammad, Mohamed, et al. "Detection of abnormal heart conditions based on characteristics of ECG signals." Measurement 125 (2018): 634-644.

[4] Wasimuddin, Muhammad, et al. "Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: A survey." IEEE Access 8 (2020): 177782-177803.

[5] Neri, Luca, et al. "Electrocardiogram monitoring wearable devices and artificial-intelligence-enabled diagnostic capabilities: a review." Sensors 23.10 (2023): 4805.

[6] Handschu, René, et al. "Telemedicine in emergency evaluation of acute stroke: interrater agreement in remote video examination with a novel multimedia system." Stroke 34.12 (2003): 2842-2846.

[7] Andrysiak, Tomasz. "Machine learning techniques applied to data analysis and anomaly detection in ECG signals." Applied Artificial Intelligence 30.6 (2016): 610-634.

[8] Abubaker, Mohammed B., and Bilal Babayiğit. "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods." IEEE transactions on artificial intelligence 4.2 (2022): 373-382.

[9] Pasha, Moghal Yaseen, et al. "Deep Learning-Based ECG Analysis for Anomaly Detection and Classification Using DCNN." 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN). IEEE, 2024.

[10] Andrysiak, Tomasz. "Machine learning techniques applied to data analysis and anomaly detection in ECG signals." Applied Artificial Intelligence 30.6 (2016): 610-634.

[11] Venkatesan, C., et al. "ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications." IEEE Access 6 (2018): 9767-9773.

[12] Kossi, KHADIDJA Ousman, et al. "Cardiovascular Disease Prediction Using Electrocardiogram (ECG) and K-Plus Nearest Neighbors Algorithm: Cases of Chadian Patients." Technium 13 (2023).

[13] Li, Taiyong, and Min Zhou. "ECG classification using wavelet packet entropy and random forests." Entropy 18.8 (2016): 285.

[14] Zaorálek, Lukáš, Jan Platoš, and Václav Snášel. "Patient-adapted and inter-patient ECG classification using neural network and gradient boosting." Neural Network World 28.3 (2018): 241-254.

[15] Warrick, Philip, and Masun Nabhan Homsi. "Cardiac arrhythmia detection from ECG combining convolutional and long short-term memory networks." 2017 Computing in Cardiology (CinC). IEEE, 2017.

[16] Al Rahhal, Mohamad M., et al. "Convolutional neural networks for electrocardiogram classification." Journal of Medical and Biological Engineering 38 (2018): 1014-1025.

[17] Moreno-Sánchez, Pedro A., et al. "ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review." Computers in Biology and Medicine (2024): 108235.

[18] Clifford, Gari D., Francisco Azuaje, and Patrick Mcsharry. "ECG statistics, noise, artifacts, and missing data." Advanced methods and tools for ECG data analysis 6.1 (2006): 18.

[19] Hong, Jianyuan, et al. "A clinical study on atrial fibrillation, premature ventricular contraction, and premature atrial contraction screening based on an ECG deep learning model." Applied Soft Computing 126 (2022): 109213.

[20] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of big data 6.1 (2019): 1-48.

[21] [1] O. Naifar, "Theoretical Framework for Tempered Fractional Gradient Descent: Application to Breast Cancer Classification," arXiv preprint arXiv:2504.18849, 2025. Available: https://arxiv.org/pdf/2504.18849v1. arxiv.org