Journal of Methodical Computing and Communications (JMCC), Vol. 2025, pp. 41-53, 2025

Comparative Evaluation of Deep Learning Models for ECG Classification Using MIT-BIH and PTB Datasets

Ahmed M.H. Darghaoth[1]*

[1] Computer Center, University of Mosul, Mosul, Iraq
***Corresponding email:** ahmed.darghaoth@uomosul.edu.iq

Received 11 July 2025**; Revised** 01 August 2025**; Accepted** 22 August 2025**; Published** 26 August 2025.

Abstract: *In MIT-BIH Arrhythmia Dataset and PTB Diagnostic ECG. Database classification of electrocardiogram (ECG) signals is essential in timely diagnosis. Electrocardiogram (ECG) analysis is one of the most widely used methods to detect cardiac abnormalities, but manual interpretation is time-consuming and prone to variability. This creates an urgent need for automated ECG classification systems that can deliver reliable results in clinical and real-time settings. In this study, three deep learning-based models, Autoencoder with Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) are compared and evaluated, trained on mixed and matched ECG data sets using MIT-BIH and PTB Diagnostic ECG datasets that have both normal and abnormal heartbeats. The evaluation base of each model is based on an evaluation of predictive accuracy, precision, recall, F1-score, training time, inference time and model size. As results show, CNN performs best in terms of classification accuracy (66.12 %), LSTM provides quite same performance but faster in inference time (65.89 %). Autoencoder with SVM model is smaller and takes less time to train; however, its overall accuracy is 59.23 %. Such results indicate the trade-off between computation speed, accuracy of diagnosis, and thus enable to specify model choice in real-time or ECG diagnostic system with limited resources.*

Keywords: ECG classification, Deep learning, Autoencoder with SVM, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Arrhythmia detection

1 Introduction

Electrocardiogram (ECG) is a very important diagnostic tool used in identification of cardiac pathology. The previous increased accessibility of good quality ECG datasets created a great motivation of creating automated classification systems that have the ability to provide timely and accurate clinical results[1], [2], [3]. More classical machine learning frameworks are often based on hand-designed features, the combination of which might not go well for other patient groups or different signal properties [4]. In contrast to deep-learning architectures, using raw ECG waveforms as an input, the directly extracted discriminative representations can aid in creating a more robust architecture and reduce features of engineering dependence on domain-specific feature[5], [6]. Current paper performs an empirical evaluation of three different deep-learning representations:

- An autoencoder-based unsupervised feature extraction method coupled with a Support Vector Machine (SVM) classifier to take advantage of compressed latent representations.
- Convolutional Neural Networks (CNNs) that are capable of automatically classifying spatially based patterns on raw waveforms.
- Long Short-Term Memory (LSTM) networks to learn relationship between heartbeat data that follows over time.

Major research goals are:

- To benchmark performance of these three different deep learning frameworks against real world ECG datasets.
- To evaluate each model against standardized Performance metrics to enable fair comparison.
- To determine generalization capability of those models in two popular ECG databases: MIT-BIH and PTB.

This paper has several contributions as follows:

- A comparative empirical analysis among autoencoder-SVM, CNN, and LSTM models in classification of ECG.
- Thorough performance analysis based on two benchmarking databases (MIT-BIH and PTB), which guarantees stability and general applicability of results.
- Shows strengths and limitations of various deep learning paradigms in ECG signal analysis which, may, inform future development of automated diagnostic systems.

This paper consists of: Section 2 literature review on ECG classification based on machine learning and deep learning methods. Section 3 describes proposed methodology and characteristics of datasets utilized in experiments. Section 4 explains training processes and structures of Autoencoder-SVM, CNN, and LSTM systems. Section 5 gives experimental results of classification, computational efficiency, and provides discussion. Lastly, Section 6 will conclude this paper with a synopsis of research findings, trade-offs of models under evaluation, as well as avenues that may be explored under future research.

2 Literature Review

Proper diagnosis of electrocardiogram (ECG) signals is central in cardiology since early diagnosis of arrhythmias and other fatal heart diseases [7], [8], [9]. With introduction of machine learning (ML) and deep learning (DL), an aspect that has brought major change to this area is automated systems, much faster and reliable compared to a manual analysis. It is a synthesis of major contributions in area, and it shows various approaches used to improve performance of classification, robustness, and clinical usefulness [8], [10], [11].

In clinical data, uncertainty in class balance is the main problem in electrocardiogram classification, due to some arrhythmic behaviors that occur much less frequently than normal heartbeats[12].

In order to address this, an ingenious combination of dedicated machine learning models, proposed by Ahamed et al. (2020) was introduced They used a technique, where loss function of an artificial neural network (ANN) is redefined by class weights, in MIT-BIH Arrhythmia and PTB Diagnostic ECG sets. Such a strategical shift allowed models to be more focused on underrepresented categories, with impressive accuracy rates of 98.06% and 97.66% on corresponding datasets. Their research highlights important role of careful loss weighting and model tuning to achieve high performance on skewed medical data which appears frequently in end-to-end clinical pipeline contexts [13]. Same year, Anaz et al. (2020) also introduced a Multi-Encryption System (MES)using Deep Autoencoder Networks (DAN/ADL). The work therefore implied usefulness of autoencoders in unsupervised feature learning and signal reconstruction. Methods capable of being incorporated along with existing ECG classification pipelines to provide resilience to signal corruption, as well as enrich recovery of the features extracted thus improving overall classification, particularly in difficult real-world case scenario [14].

To have a basic idea about recurrent neural networks in this area, a survey by Lindemann et al. (2021) on Long Short-Term Memory (LSTM) networks to forecast time series is recommended Their efforts include a description of known derivatives of LSTM cell and LSTM network architectures and divide them into more than one category: optimized or interacting cell states. They also negate important considerations of proper time series prediction, such as short-term and long-term memory behavior, multimodal and multi-step ahead forecasts, and errors propagation. In survey, they arrive at conclusion that sequence-to-sequence networks with partial conditioning are preferable to solve these requirements over other LSTM architectures, which substantiates theoretical propositions on applicability of LSTM in time-variant bio-signals such as ECG [15].

Signal quality is a very important aspect in analysis of an ECG, noise may be a problem that seriously compromises classification accuracy. To resolve this conundrum, Errabih et al. (2022) developed a more optimized 12-layer CNN model to classify arrhythmia and used MIT-BIH Arrhythmia dataset in study. The main difference between their work and that of others is that modern CNNs are considered as a solution to this problem, because along with proposed model, wavelet-based

was used to depend on signal and eliminate noise before feeding data to CNN. Moreover, their project presented a self-regulated strategy of feature learning and effective optimization. Due to this innovation, an online retraining may take less time when involvement of experts is minimal, it improves efficiency of system as it is quick to be deployed in real-time. This given system is suitable to classify five micro-classes of heartbeat with competitive level of accuracy and enhanced efficiency of operation[16]. More recently, Sattar et al. (2024) proposed a new pipeline which embedded ECG images of clinical centers around Pakistan into time-series signals. In some different deep learning architectures, signals have been injected into them, such as Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and Self-Supervised Learning Encoders (SSL). Their CNN model was predicted to be slightly more than 92%, which showed a viability of an image to signal transformation approach. It can be mixed with deep learning, and it is a potentially feasible path towards achieving a real-time ECG monitoring whose output does not need extensive human annotation and expert preprocessing to make it ready to be used in clinic [17]. In the same year, Zabihi et al. (2024) devised one such system, possessing two supplementary subsystems, first relied on ResNet-LSTM deep learning architecture, which excelled at pattern learning on raw signals. The second exploited combination of handcrafted features, a Random Forest classifier and Synthetic Minority Over-sampling Technique (SMOTE) to balance class imbalance. Without these two systems being combined, an excellent accuracy of 99.26% on MIT-BIH dataset was obtained. This paper gives strong arguments to the idea that combinational usage of learned features of deep networks and carefully designed features may give a substantial performance boost on complex clinical classification tasks [18].

Issue of data imbalance in ECG dataset is widespread because some arrhythmias are more infrequent than others. This literature review presented shows that some effective measures are adopted to reduce impact of such a problem, making sure that key but not well-represented types of arrhythmias are correctly identified. These are loss weighting [13], signal segmentation and digitization[17], oversampling using SMOTE [18], automatic feature learning and denoising [16], and autoencoder latent compression [11], [12], [19]. All of these ways aim at guarantee that all types of arrhythmias, the most important yet underrepresented ones are detected reliably and bring us closer to the idea of clinically feasible automated analysis systems working with ECG.

3 Proposed Method

This paper provides a comparative discussion of three deep learning models of automatic classification of the ECG signal Autoencoder in combination with Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). To conduct experiments the merged dataset with two well-recognized databases, namely, MIT-BIH Arrhythmia Database and PTB Diagnostic ECG Database. All models were trained and tested in the same conditions to compare their efficiency along some dimensions, such as accuracy, training time, inference speed, and memory size.

3.1 ECG Heartbeat Database

This paper utilizes two openly accessible and broadly confirmed electrocardiogram (ECG) datasets, namely, MIT-BIH Arrhythmia Database and PTB Diagnostic ECG Database. They were chosen because of their broad range of datasets, as they are well established, and rather common in domain of cardiac signal analysis research.

3.1.1 MIT-BIH Arrhythmia Dataset

A benchmark dataset is MIT-BIH Arrhythmia Database , a collection maintained by the PhysioNet collection [20]. Dataset includes 48 half hours of ECG data of 47 different patients recorded. All recordings were digitized at 360Hz sampling frequency, and labeled by cardiologists according to beat-level labels using AAMI EC57 standard. The data covers a broad range of variants of arrhythmia, makes it applicable and suitable to construct strong algorithms in heart beat classification. Given this dataset, the preprocessed form of it, available as mitbih_train and mitbih_test is used in this work, each sample containing 187-time steps of segmented heartbeats and correct labels (0 to 4) indicating classes. The classes equate to standard beats and that of four types of arrhythmias.

3.1.2 PTB Diagnostic ECG Database

PTB Diagnostic ECG database collected by Federal Institute of Physics and Technology (PTB) in Germany and found on PhysioNet has also been uploaded[21]. It has 549 ECG recordings of 290 healthy and also sick (myocardial infarction, bundle branch block) subjects. All records contain 15 signals that have been simultaneously captured and sampled at 1000 Hz. In this work, two selected datasets will be employed that are named ptb_normal and ptbdb_abnormal, which refer to normal and abnormal heartbeats respectively. Samples were coded as either normal (label 0) or not (label 1), each of them is a 187-point sample. Such subsets allow increasing dataset diversity and enhance generalization capacity of model to other population groups.

3.1.3 Dataset Integration

Combination of MIT-BIH and PTBDB datasets allows constructing a single training and evaluation pipeline. mitbih_train and ptbdb_normal are merged together to form training set, same is done with mitbih_test and ptbdb_abnormal to form test set. With such a setup, training set will contain both arrhythmic and normal patterns with test set containing heterogeneous, unseen abnormalities.

The following preprocessing procedure was done:

- **Handling Missing Values:**
  The two data sets are validated and curated by PhysioNet. However, missing or corrupted entries in all ECG records were scanned. Incomplete time steps were discarded as a record. The proportion of removed samples was miniscule (<0.5%), to guarantee that there will be no serious data loss or introduction of bias.
- **Normalization:**
  Since ECG waveforms need to be scale-invariant across patients and datasets, all ECG signal inputs were down-sampled to a common number of 187 time steps. Then, each segment of the ECG was normalized according to Z-scoring:

$$x_i' = \frac{\mu - x_i}{\sigma} \tag{1}$$

where x_i is actual signal value, mean and standard deviation of this segment is described by μ and σ. This transformation ensures that mean of each signal is zero, variance is one and thus any variation in amplitude does not have any influence on model training.

- **Label and Format Consistency:**
  Nominal number of classes in the provided MIT-BIH dataset is five (normal and four arrhythmias), whereas PTB dataset is assigned with two categories (normal and abnormal). To standardize label coding
- **MIT-BIH classes were coded as integers 0-4.**
- **PTB classes were coded as normal as 0 and as abnormal as 1.**

We standardized all samples in a similar format 187 x 1 time-series vector and integer labeling. This guaranteed preserved label quality of merged dataset with exact integrity of data input to deep learning models. Preprocessing pipeline ensured compatibility of two datasets, and increased robustness by simulating heterogeneity in ECG recordings in real world.

3.2 Model Training

Training of model was performed on a host computer with Intel Core i5-1135G7 CPU, 11th generation and 8 GB of RAM Matlab 2024A. Adam optimization is used to train each model, with changes in mini-batch size, epochs, and learning rate. Three different deep neural networks are trained, each of which has its own architecture and training conditions.

3.3 Autoencoder with SVM

In first approach, a hybrid model combining an Autoencoder and Support Vector Machine (SVM) is employed for ECG classification. The Autoencoder is designed to learn compressed representations of raw ECG signals through unsupervised reconstruction. The encoder comprises three fully connected layers with 128, 64, and 64 neurons, respectively, interleaved with ReLU activations and a dropout layer (rate = 0.2) to mitigate overfitting. The final 64-dimensional bottleneck layer serves as latent feature space. Decoder symmetrically reconstructs original input using mirrored fully connected layers and is optimized using a mean squared error loss function.

The model was trained using Adam optimizer with a learning rate of 0.0001, a mini-batch size of 32, and for a total of 20 epochs. Training data were shuffled at every epoch to ensure generalization, and training progress was monitored via real-time loss visualization. Once Autoencoder was trained, it was used to extract 64-dimensional latent features from both training and testing datasets. These features were normalized using Z-score normalization and classified using a multi-class SVM with a radial basis function (RBF) kernel. SVM was implemented via an error-correcting output code (ECOC) framework to handle multi-class nature of ECG categories. This two-stage pipeline effectively combines unsupervised representation learning with supervised classification, providing a computationally lightweight yet structured model for ECG signal interpretation.

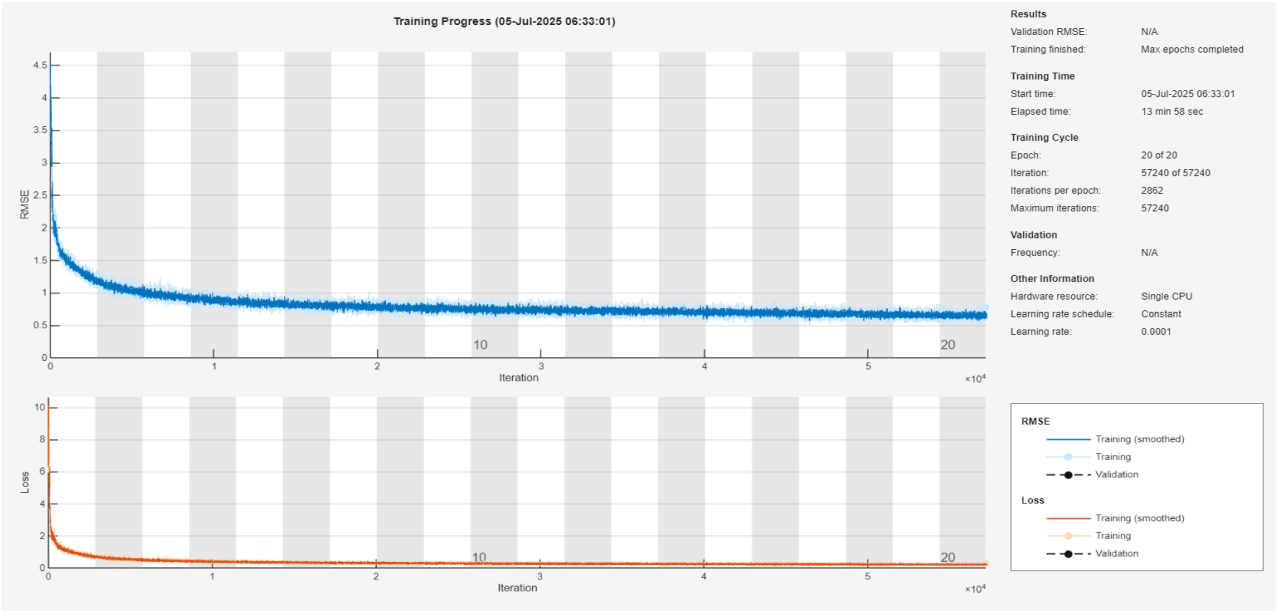

Figure 1: Autoencoder with SVM model training progress.

During training, model showed a gradual reduction in reconstruction loss, indicating effective feature learning, as shown in Figure 1. Experimental protocol was constructed with intention of gauging cpu-time consumed by process of training regression model and estimation of associated classification accuracy. Training period took 13 minutes and 58 seconds depicting positive trade-off between temporal efficiency and computational resources. Reduction in classification achieved by other models was rather large but because we consider parsimonious architecture of this model and limited training requirements, this model is highly relevant to systems characterized by constraints associated with computation.

3.4 Convolutional Neural Network (CNN)

Current paper assesses another approach, classification of ECG signals by using of a one-dimensional CNN built in a way that identifies spatial patterns directly in raw inputs. Each of ECG traces was originally a 1 x 187 vector, but it was divided into [187 x 1 x 1] in 2 dimensions, to comply with input requirements of CNN. Architecture was based on use of two convolutional blocks that used depth-separable convolutions, coming after data input layer. The first convolutional layer applies 64 filters of size [3×1], followed by batch normalization, ReLU activation, and max pooling with a [2×1] window. The second block uses 128 filters of the same size, again followed by batch normalization, ReLU, and max pooling. A dropout layer with a rate of 0.5 is included after the second pooling layer to reduce overfitting. Finally, a fully connected layer maps features number of ECG classes, followed by a softmax layer and a classification output.

The network was trained using Adam optimizer with a learning rate of 0.001, a mini-batch size of 32, and for 20 epochs. Data shuffling was enabled at every epoch to improve generalization. Training process was monitored through dynamic visualization of loss and accuracy curves. Z-score normalization was applied to each ECG sample to ensure consistent feature scaling across dataset. This model exhibited rapid convergence during training, with steadily increase in accuracy and decrease in loss. Training was performed over a span of approximately 109 minutes, as shown in

Figure *2*,dupartially because depth and parameter complexity of network. CNN surpassed other models in classification accuracy, indicating its robustness in learning spatial patterns from ECG data.

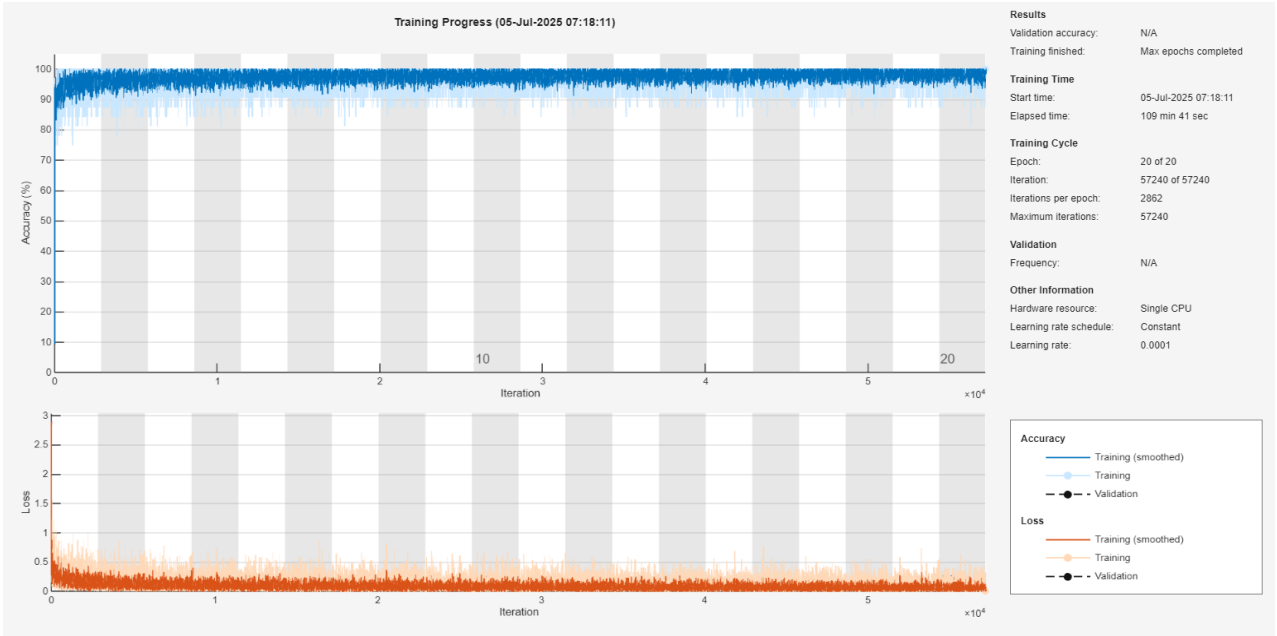

Figure 2: CNN model training progress

3.5 Long Short-Term Memory (LSTM)

An LSTM network was implemented to capture the temporal dependencies within ECG time-series data. The model accepts input sequences of length 187 via a sequence Input Layer. Architecture includes two stacked LSTM layers: first with 128 hidden units and output mode set to "sequence", enabling it to pass entire time-series to the next layer; second with 64 hidden units and output mode "last" to extract final time step's context vector. Each LSTM layer is followed by a dropout layer with a dropout rate of 0.3 to mitigate overfitting. A fully connected layer maps the temporal features to number of output classes, followed by a softmax activation and a classification layer to perform multiclass heartbeat classification.

Training was conducted using Adam optimizer with a learning rate of 0.0001, a mini-batch size of 32, and for 20 epochs. Epoch-wise shuffling was applied to improve robustness. Training progress was monitored using real-time plots of accuracy and loss to ensure convergence stability and detect signs of overfitting. Training performance demonstrated consistent improvement in accuracy and a reduction in training loss over time, as shown in

Figure 3. when an LSTM neural network was applied to classification of human activities, network trained in about 15 minutes and inferred in only a few seconds. This kind of performance shows that LSTM models can be useful in real-time monitoring scenarios that require quick action in response to monitored system besides a high classification accuracy.

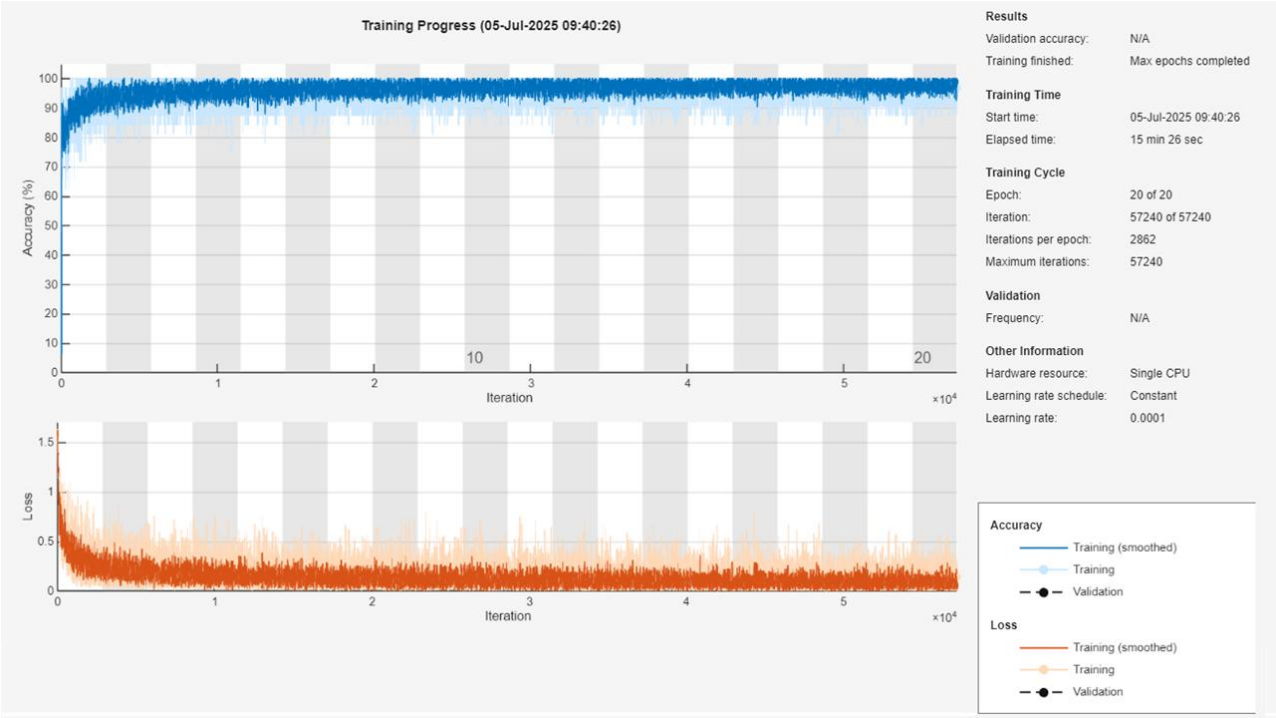

Figure 3: LSTM model training progress

3.6 Training Behavior and Comparative Summary

A detailed visualization of progress of training of each model has been done by analysis of accuracy and loss curves. The autoencoder showed consistent convergence of loss in reconstructed samples; the CNN reached higher training accuracy since beginning; LSTM demonstrated steady performance with low loss a gradual and steady increase of accuracy. Collectively, all these findings indicate that process of optimization and learning followed predictable trends in each model.

4 Result and Discussion

The following section will provide a stringent point-by-point analysis of three different machine learning structures which entail Autoencoder with SVM, CNN, and LSTM network. This is compared in key performance indicator such as train accuracy, precision, recall, F1-score and computation speed, measured in terms of training time, inference time and model size. These metrics perform a pivotal role in determining practical feasibility and clinical benefit of automated electrocardiogram signal classification systems.

4.1 Classification Performance Metrics

Accuracy, Precision, recall, and F1-Score are the major measures of the classification task performance.

4.1.1 Accuracy

By comparing , Figure 4: Model performance comparison, overall classification accuracy of CNN (66.12%) LSTM (65.89%) is very much better than that of Autoencoder with SVM (59.23%). Results prove that deep learning architecture is better at finding discriminative, complex features in ECG signals.
Accuracy measures the overall proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

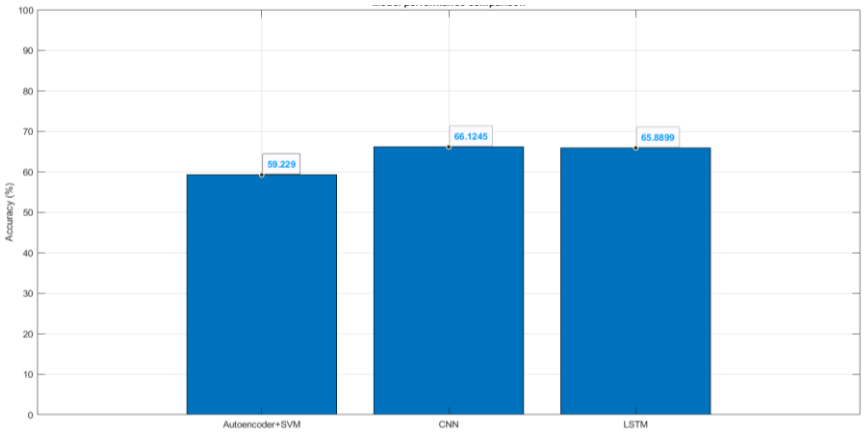

Figure 4: Model performance comparison

4.1.2 Precision

Model Precision Average comparison figure, as shown in Figure 5, highlights CNN (71.54%) scoring highest average precision, followed by LSTM (68.63%). The Autoencoder with SVM (35.85%) exhibits remarkably low precision. High precision is vital in medical diagnostics to minimize false positives, thereby reducing unnecessary interventions and patient burden. Precision quantifies proportion of correctly predicted positive samples among all predicted positives:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

where TP denotes the number of true positives, and FP denotes the number of false positives.

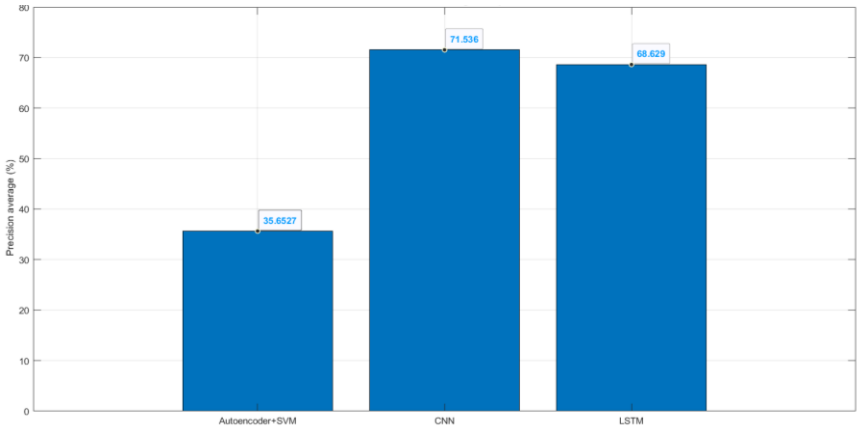

Figure 5: Model Precision Average comparison

4.1.3 Recall

Conversely, Figure 6 shows Autoencoder with SVM (87.10%) achieving highest average recall, surpassing CNN (80.40%) and LSTM (77.34%). High recall is critical to minimize false negatives, a paramount in detecting life-threatening arrhythmias, even at the potential expense of increased false positives. Recall measures proportion of correctly predicted positive samples among all actual positives:

$$Reccall = \frac{TP}{TP + FN} \quad (4)$$

where TP, and FN denote the number of true positives, and false negatives, respectively.

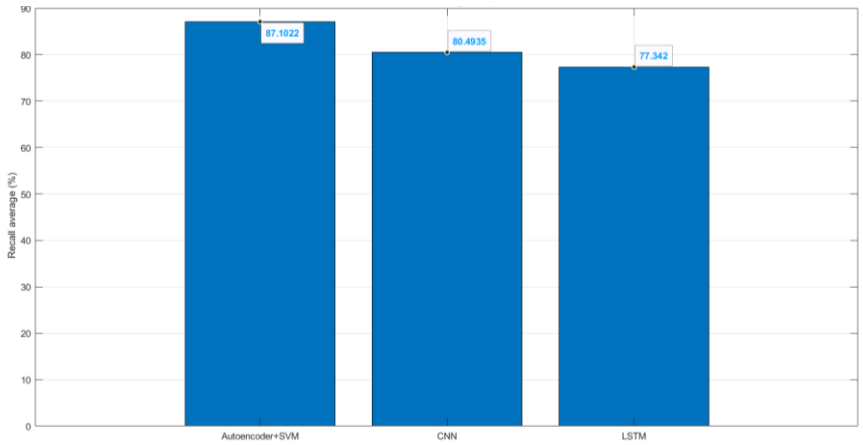

Figure 6: Model Recall Average comparison

4.1.4 F1-Score

The F1-Score, a harmonic mean of precision and recall, provides a balanced measure of performance, especially crucial for imbalanced datasets prevalent in medical applications. Model F1-Score Average comparison, as shown in Figure 7 indicates that CNN (66.75%) offers the best balance between precision and recall, followed by LSTM (62.96%). The Autoencoder with SVM (39.21%) performs poorly in this integrated metric, suggesting an imbalance in its precision-recall trade-off. The F1-score is the harmonic mean of precision and recall, providing a balanced metric especially for imbalanced datasets:

$$F1 = \mathbf{2}.\frac{Precision\,.Recall}{Precision + Recall} \tag{5}$$

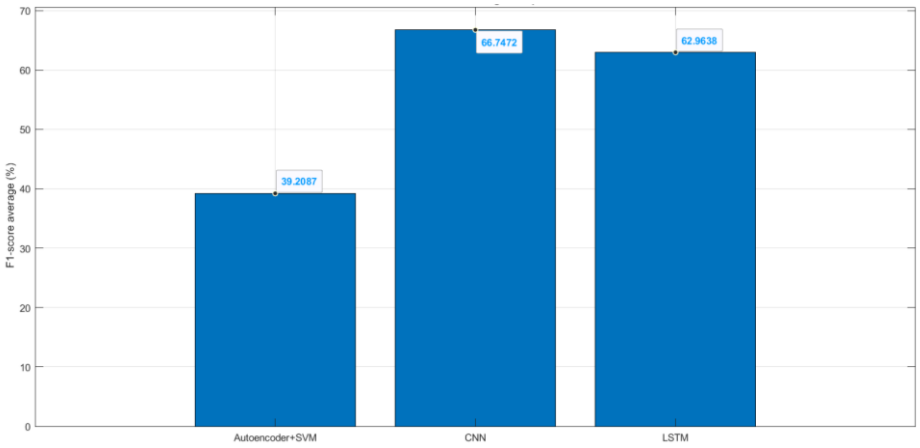

Figure 7: Model F1-Score Average comparison

Collectively, these metrics reveal that while Autoencoder with SVM excels in recall, its significantly lower precision and F1-score indicate a propensity for false alarms. CNN, in contrast, consistently provides the most balanced and robust classification performance. Comparative summary of classification performance of models is provided in Table 1.

Table 1. A comparative summary of classification performance metrics of models.

Model	Accuracy (%)	F1-Score Average (%)	Precision Average (%)	Recall Average (%)
Autoencoder with SVM	59.23	39.20	35.65	87.10
CNN	66.12	66.74	71.53	80.49
LSTM	65.89	62.96	68.629	77.34

4.2 Computational Efficiency Metrics

Efficiency of model deployment is assessed through training time, inference time, and memory usage:

4.2.1 Inference Time

As presented in model inference time comparison, shown in Figure 8, inference time for all models was measured under identical hardware and software conditions (Intel Core i5-1135G7 CPU, 8 GB RAM, MATLAB 2024a), and the same dataset size, ensuring fairness in comparison. Test set consisted of 21,870 heartbeat samples from merged MIT-BIH and PTB databases. Inference time was defined as total elapsed time required to generate predictions for the complete test set. For the Auto-encoder + SVM pipeline, reported inference time of 157.07 seconds appears disproportionately high compared to CNN (15.43 s) and LSTM (3.61 s), also relative to its training time (859.20 s). This discrepancy arises from two main factors:

- Two-stage process: inference involved (a) passing each sample through encoder to obtain a 64-dimensional latent vector, and (b) classifying these vectors using a multi-class SVM with an RBF kernel implemented through an Error-Correcting Output Codes (ECOC) framework.
- Implementation constraints: SVM classifier in MATLAB processes samples in a per-sample execution mode rather than optimized batch mode. Thus, kernel evaluations were performed repeatedly for each of the 21,870 test samples. Computational complexity of RBF-SVM prediction is approximately, $N_{test} \times N_{SV}$,where N_{test} is number of test samples and N_{SV} is number of support vectors. This leads to considerably longer inference times compared to CNN and LSTM models, which leverage parallelized matrix multiplications and batch execution.

Hence, long inference time does not reflect inefficiency in Autoencoder but rather computational burden of kernel-based SVM stage in large-scale testing. Future optimization strategies include replacing RBF kernel with a Linear SVM for faster prediction, applying support vector reduction techniques, or using kernel approximations such as Random Fourier Features. These approaches could substantially reduce inference time while maintaining acceptable classification accuracy, improving real-time applicability of Autoencoder with SVM pipeline.

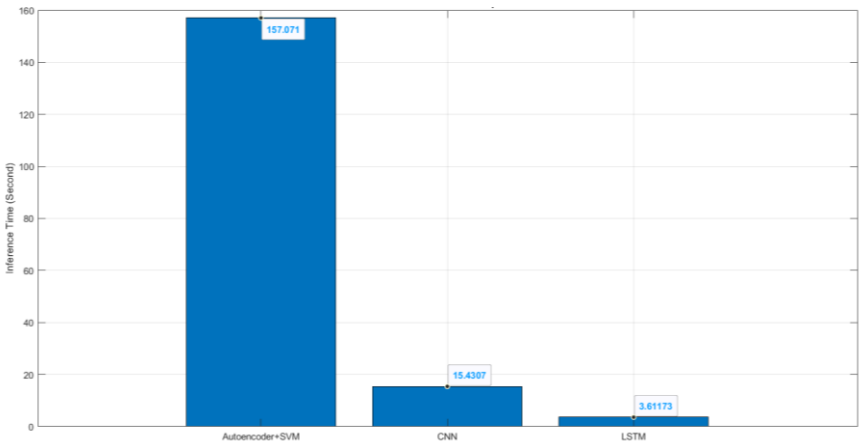

Figure 8: Model inference time comparison

4.2.2 Training Time

Figure 9 illustrates that CNN (6805.32 seconds) requires longest training duration. Both LSTM (938.32 seconds) and Autoencoder with SVM (859.20 seconds) offer significantly faster training times, which is beneficial for rapid model iteration and development cycles.

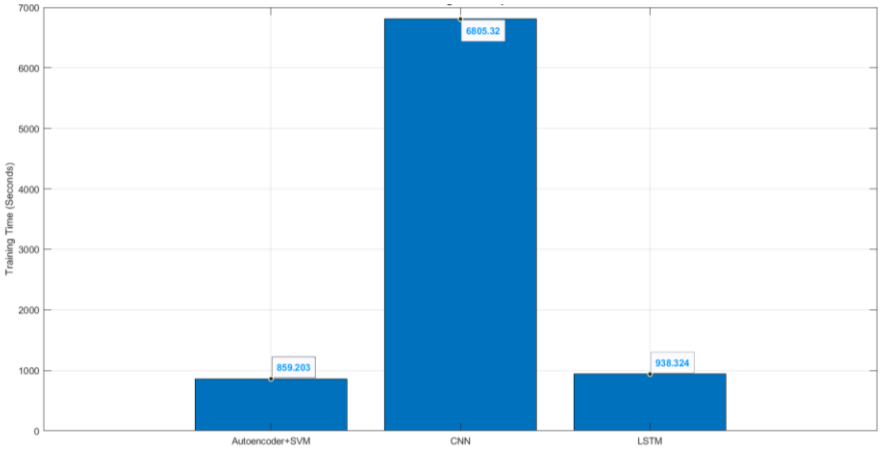

Figure 9: Model training time comparison

4.2.3 Model Memory Size

Model memory size indicates that Autoencoder with SVM (0.305 MByte) has the smallest memory size, followed by CNN (0.580 MByte), while LSTM (0.826 MByte) requires most memory, as shown Figure 10. A comparative summary of Computational Efficiency of Models is provided inTable 2.

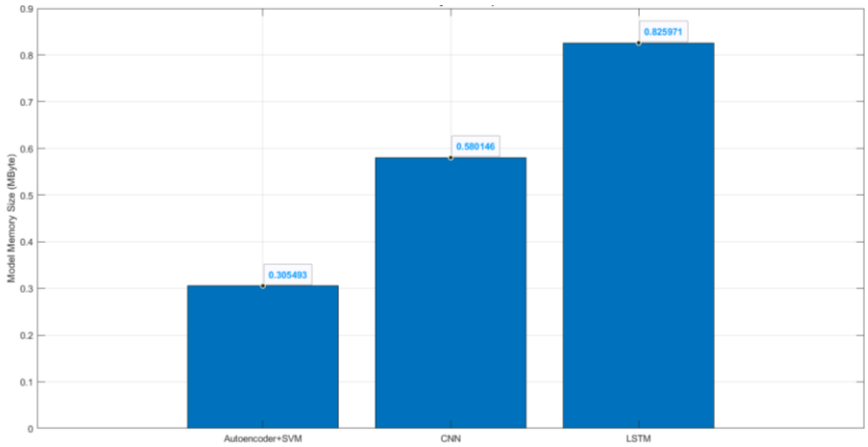

Figure 10: Model memory size comparison

Table.2. Comparative Computational Efficiency of Models

Model	Training Time (s)	Inference Time (s)	Model Size (MB)
Autoencoder with SVM	859.20	157.07	0.31
CNN	6805.32	15.43	0.58
LSTM	938.32	3.61	0.83

Analyzing computational Efficiency of Models, LSTM model demonstrates the most advantageous profile for real-time applications, exhibiting significantly faster inference times (3.61 s) compared to CNN (15.43 s) and Autoencoder with SVM (157.07 s), despite having the largest memory footprint (0.83 MB). While the Autoencoder with SVM is lightweight (0.31 MB) and fast to train, its impractical inference duration severely limits its utility. CNN, though robust in classification performance, demands substantial training resources. Therefore, for scenarios prioritizing immediate diagnostic feedback and operational responsiveness in ECG analysis, LSTM offers superior overall practical benefit by balancing acceptable training cost with highly efficient real-time operation.

4.3 Confusion Matrices

Confusion matrices for three models offer a granular view of per-class classification performance, highlighting specific strengths and weaknesses. Assuming classes 1 through 5 represent distinct ECG beat types:

4.3.1 Autoencoder with SVM Model

Exhibits a pronounced bias towards Class 1 (18117 correctly classified), as shown in Figure 11. However, this comes at a significant cost: a very large proportion of instances from minority classes (Class 2: 10902, Class 3: 1099, Class 4: 124, Class 5: 1069) are incorrectly misclassified as Class 1. This pervasive misclassification explains its high overall recall (as many true positives for Class 1 are correctly identified, and many others are "recalled" as Class 1) but also its severely low precision and F1-score, making it unreliable for precise multi-class differentiation.

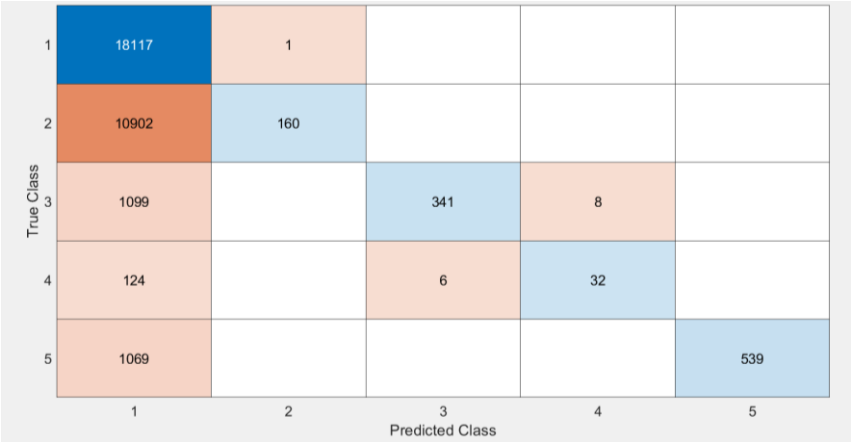

Figure 11: Autoencoder with SVM Model Confusion Matrix

4.3.2 CNN Model

Strong performance on Class 1 (18068 correctly classified), robust performance on Classes 3 (1299) and 5 (1557), as shown in Figure 12. The primary challenge lies in misclassifying a substantial number of Class 2 instances (10176) as Class 1, indicating difficulty in distinguishing between these two specific categories. This contributes to a higher false negative rate for Class 2. Overall, diagonal dominance across multiple classes reflects its strong accuracy and balanced F1-score.

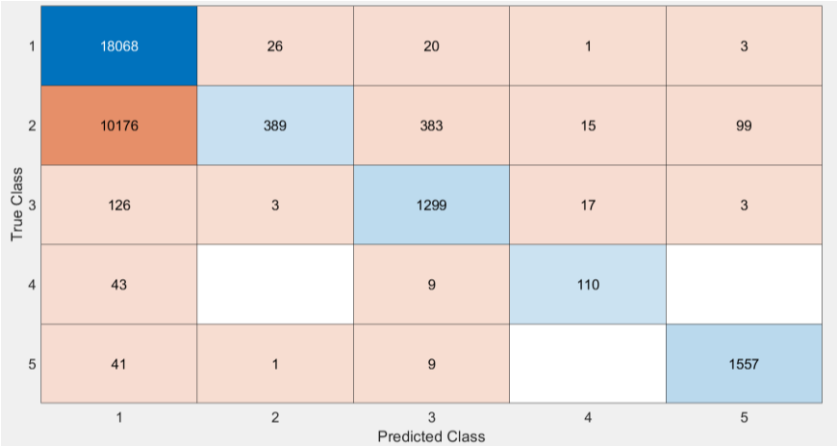

Figure 12: CNN Model Confusion Matrix

4.3.3 LSTM Model

Similar to CNN, LSTM demonstrates strong performance on Class 1 (18052 correctly classified), effective classification of Classes 3 (1308) and 4 (88). It also faces challenges with misclassifying Class 2 instances (9787) as Class 1. A notable interaction is observed between Class 2 and Class 5, with 498 Class 2 instances predicted as Class 5, and 498 Class 5 instances predicted as Class 2), as shown in Figure 13. This suggests potential feature overlap or ambiguity between these two classes that LSTM struggles to resolve.

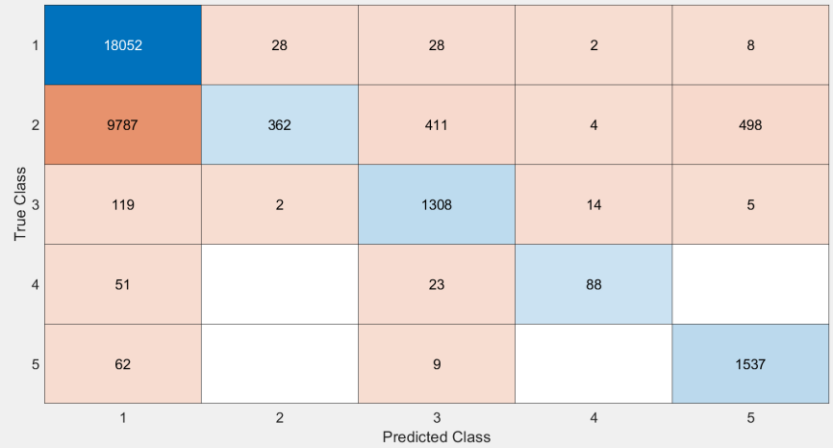

Figure 13: LSTM Model Confusion Matrix

5 Conclusion

This study rigorously evaluated performance of Autoencoder with SVM, CNN, and LSTM models for electrocardiogram (ECG) signal classification across key metrics including classification accuracy, precision, recall, F1-score, and computational efficiency (training time, inference time, and memory size). Findings consistently demonstrate that deep learning architectures, specifically CNN and LSTM, significantly outperform Autoencoder with SVM model in overall classification efficacy. CNN model emerged as the most robust classifier, achieving highest balanced performance across accuracy (66.12%), precision (71.54%), and F1-score (66.75%). This indicates its superior ability to accurately identify various ECG beat types while minimizing both false positives and false negatives, making it highly suitable for applications where diagnostic reliability is paramount. Conversely, LSTM model proved exceptionally efficient in real-time inference (3.62 seconds), significantly outperforming both CNN (15.43 seconds) and Autoencoder with SVM (157.07 seconds). This speed, coupled with competitive classification performance (Accuracy: 65.89%, F1-Score: 62.96%), puts LSTM as an ideal candidate for applications requiring immediate diagnostic feedback and rapid processing. Auto-encoder with SVM model, while exhibiting fastest training time and smallest memory size, suffered from critically low precision (35.85%) and an impractical inference time. Its pronounced bias towards majority class, as shown by confusion matrices, severely limits its utility for comprehensive multi-class arrhythmia detection in clinical settings. In essence, selection of an optimal model for automated ECG classification is a trade-off between diagnostic accuracy and operational efficiency. For applications demanding highest classification robustness and reliability, CNN is recommended. For real-time and resource-sensitive deployments where rapid inference is critical, LSTM offers a compelling solution. Future work should explore hybrid deep learning strategies or advanced regularization techniques to further enhance inter-class discrimination, particularly for challenging minority classes, thereby improving clinical translatability of these automated systems.

ACKNOWLEDGMENT

Deepest thanks are due to Computer Center at University of Mosul, Iraq, for their positive support**.**

REFERENCES

[1] Z. Kotsialou, N. Makris, and S. Gall, “Fundamentals of the electrocardiogram and common cardiac arrhythmias,” *Anaesth. Intensive Care Med.*, vol. 25, no. 3, pp. 219–222, Mar. 2024, doi: 10.1016/j.mpaic.2023.11.014.

[2] M. Ahmad, A. Ahmed, H. Hashim, M. Farsi, and N. Mahmoud, “Enhancing Heart Disease Diagnosis Using ECG Signal Reconstruction and Deep Transfer Learning Classification with Optional SVM Integration,” *Diagnostics*, vol. 15, no. 12, p. 1501, Jun. 2025, doi: 10.3390/diagnostics15121501.

[3] P. A. Moreno-Sánchez *et al.*, “ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review,” *Comput. Biol. Med.*, vol. 172, p. 108235, Apr. 2024, doi: 10.1016/j.compbiomed.2024.108235.

[4] R. Ranjan, B. C. Sahana, and A. K. Bhandari, “Deep Learning Models for Diagnosis of Schizophrenia Using EEG Signals: Emerging Trends, Challenges, and Prospects,” *Arch. Comput. Methods Eng.*, vol. 31, no. 4, pp. 2345–2384, May 2024, doi: 10.1007/s11831-023-10047-6.

[5] 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, Sep. 2023, doi: 10.3389/fphys.2023.1246746.

[6] M. Khalid, C. Pluempitiwiriyawej, S. Wangsiripitak, G. Murtaza, and A. A. Abdulkadhem, “The Applications of Deep Learning in ECG Classification for Disease Diagnosis: A Systematic Review and Meta-Data Analysis,” *Eng. J.*, vol. 28, no. 8, pp. 45–77, Aug. 2024, doi: 10.4186/ej.2024.28.8.45.

[7] N. Katal, S. Gupta, P. Verma, and B. Sharma, “Deep-Learning-Based Arrhythmia Detection Using ECG Signals: A Comparative Study and Performance Evaluation,” *Diagnostics*, vol. 13, no. 24, p. 3605, Dec. 2023, doi: 10.3390/diagnostics13243605.
[8] D. Patil, N. L. Rane, P. Desai, and J. Rane, “Machine learning and deep learning: Methods, techniques, applications, challenges, and future research opportunities,” in *Trustworthy Artificial Intelligence in Industry and Society*, Deep Science Publishing, 2024. doi: 10.70593/978-81-981367-4-9_2.
[9] P. K. Singh, S. Akhtar, A. Gupta, and S. Singh, “The Clinical Relevance of ECG Parameters in the Prediction of Cardiac Mortality: A Comprehensive Review,” *Open Bioinforma. J.*, vol. 17, no. 1, Jun. 2024, doi: 10.2174/0118750362295563240620111209.
[10] S. Asif *et al.*, “Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision,” *Arch. Comput. Methods Eng.*, vol. 32, no. 2, pp. 853–883, Mar. 2025, doi: 10.1007/s11831-024-10148-w.
[11] O. Shobayo and R. Saatchi, “Developments in Deep Learning Artificial Neural Network Techniques for Medical Image Analysis and Interpretation.,” *Diagnostics (Basel, Switzerland)*, vol. 15, no. 9, Apr. 2025, doi: 10.3390/diagnostics15091072.
[12] A. Shah, D. Singh, H. G. Mohamed, S. Bharany, A. U. Rehman, and S. Hussen, “Electrocardiogram analysis for cardiac arrhythmia classification and prediction through self attention based auto encoder,” *Sci. Rep.*, vol. 15, no. 1, p. 9230, Mar. 2025, doi: 10.1038/s41598-025-93906-5.
[13] M. A. Ahamed, K. A. Hasan, K. F. Monowar, N. Mashnoor, and M. A. Hossain, “ECG Heartbeat Classification Using Ensemble of Efficient Machine Learning Approaches on Imbalanced Datasets,” in *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, IEEE, Nov. 2020, pp. 140–145. doi: 10.1109/ICAICT51780.2020.9333534.
[14] A. Sameer Anaz Raid Rafi Omar Al-Nima Moatasem Yaseen Al-Ridha, “Multi-Encryptions System Based on Autoencoder Deep Learning Network.”
[15] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, “A survey on long short-term memory networks for time series prediction,” in *Procedia CIRP*, Elsevier B.V., 2021, pp. 650–655. doi: 10.1016/j.procir.2021.03.088.
[16] E. Abdelhafid *et al.*, “ECG Arrhythmia Classification Using Convolutional Neural Network,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 12, no. 7, pp. 186–195, Jul. 2022, doi: 10.46338/ijetae0722_19.
[17] S. Sattar *et al.*, “Cardiac Arrhythmia Classification Using Advanced Deep Learning Techniques on Digitized ECG Datasets,” *Sensors*, vol. 24, no. 8, Apr. 2024, doi: 10.3390/s24082484.
[18] F. Zabihi, F. Safara, and B. Ahadzadeh, “An electrocardiogram signal classification using a hybrid machine learning and deep learning approach,” *Healthc. Anal.*, vol. 6, Dec. 2024, doi: 10.1016/j.health.2024.100366.
[19] A. S. Anaz, M. Y. Al-Ridha, and R. R. O. Al-Nima, “Signal multiple encodings by using autoencoder deep learning,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 1, pp. 435–440, Feb. 2023, doi: 10.11591/eei.v12i1.4229.
[20] 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, doi: 10.1109/51.932724.
[21] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet,” *Circulation*, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.