Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 5, October 2026, pp. 4325∼4337
ISSN: 2302-9285, DOI: 10.11591/eei.v15i5.11515

CNN-BiLSTM based model for heart disease classification using deep learning techniques

Mohammad Azhar[1], Lazer Mary Gladence[2]
[1]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
[2]School of Computing, Sathyabama Institute of Science and Technology, Chennai, India

Article Info

Article history:

Received Sep 25, 2025
Revised Aug 26, 2026
Accepted Aug 28, 2026

Keywords:

Bidirectional long short-term memory
Convolutional neural network-long short-term memory
Deep learning
Heart disease
Machine learning

ABSTRACT

Heart disease remains a main cause of death worldwide, showing the necessity for accurate and timely automated diagnostic systems. Hence, precise and early detection is necessary and is done by some existing machine learning (ML) and deep learning (DL) algorithms. Existing ML and DL models, like convolutional neural network (CNN), long short-term memory (LSTM), CNN-LSTM, and some hybrid approaches, shows good performance. However, they struggle to capture sequential dependencies and spatial feature interactions simultaneously. Furthermore, hybrid CNN–BiLSTM models to tabular heart disease datasets remains unexplored. To address this gap, this paper, proposes a convolutional neural network–bidirectional long short-term memory (CNN-BiLSTM) hybrid model. This proposed model employs bidirectional processing to improve the temporal features extraction. The proposed hybrid model accurately stores sequential data of the past and the future to do predictions. The experimental validation is done with two different heart disease datasets. The implementation includes learning rate scheduling and hyperparameter adjustment for optimisation. Experimental results on two benchmark datasets demonstrate the proposed model effectiveness. The proposed model got 98.54% accuracy with heart disease dataset and 81.94% accuracy with heart failure clinical dataset, done using 5-fold cross-validation.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Lazer Mary Gladence
School of Computing, Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu, India
Email: lgladence@gmail.com

1. INTRODUCTION

Heart disease is a serious health issue nowadays, requiring efficient diagnostic tools for early detection and treatment. Existing models rely on clinical assessments and manual interpretations, which are time-consuming and creates errors. The usage of deep learning (DL) has revolutionized the healthcare domain, providing automated and accurate disease detection mechanisms [1]. Also, it takes more time well as is prone to errors when done manually. To improve diagnosis and therapy, DL models should be utilized for early automated detection.

Although recent DL approaches have significantly improved heart disease prediction (HDP), many existing models rely on shallow machine learning (ML) techniques or single DL architectures. Such approaches often struggle to capture both sequential dependencies and spatial feature interactions present in structured clinical data effectively, thus limiting diagnostic performance.

Some of the existing works given for heart disease detection are listed here. Some of the existing works use a DL model that mixes bidirectional long short-term memory (BiLSTM) with convolutional neural network (CNN) [2]. A cluster-based BiLSTM (C-BiLSTM) was given in [3]. Studies are done to see how well DL models can find valvular heart diseases (VHD) using phonocardiography (PCG) recordings [4]. Different long short-term memory (LSTM) based DL model versions are suggested for classification purposes [5]–[7]. The BiLSTM DL model correctly sorts PVC beats, APC beats, and PB beats into three groups [8]. Consequently, another work develops the effective adaptive ensembled deep CNN-BiLSTM classifier model based on whale-marine optimization [9]. Another study looked all three types of DL algorithms extended DL (ETDL), integrated DL, and hybrid DL that employ DL with other models to predict HD [10]. Another research looks at some of the ways in artificial intelligence (AI) which is already enhancing the treatment of heart disease [11]. A lot of interesting research has been done on how to use recurrent models like LSTM and CNNs to sort heart diseases into groups. Still, there are obstacles to overcome when dealing with sequential dependencies and trying to successfully extract temporal patterns.

Despite improvements in ML and DL, the existing researches primarily rely on shallow learning methods or single-model architectures. These existing approaches do not consider capturing of both sequential dependencies and spatial feature interactions effectively. Furthermore, only few works have explored hybrid CNN–BiLSTM architectures for medical tabular data. Thus, this study works with this gap by proposing a hybrid DL model integrating convolutional feature extraction with bidirectional temporal modeling.

Unlike existing works that primarily focus on ECG signals, wearable IoT data streams, phonocardiogram recordings, or electronic health records (EHR), this study applies a hybrid CNN–BiLSTM architecture to structured clinical heart disease datasets. By combining convolutional feature extraction with bidirectional temporal learning, the proposed model validates higher classification accuracy, recall, and robustness across two datasets and comparison is done with existing CNN, LSTM, and CNN-LSTM models. This work main contributions are given below:

- An innovative architecture is given for DL that exploits bidirectional processing to improve feature extraction from ECG and heart disease datasets termed as hybrid CNN-BiLSTM.
- The proposed algorithm is compared extensively against CNN, LSTM, and CNN-LSTM models, showing that it achieves better classification accuracy.
- Applying scheduling for learning rates and sophisticated hyperparameter adjustment to boost performance.
- To demonstrate robustness and generalization capability the proposed model is validated on two open-source heart disease datasets using 5-fold cross-validation.

The rest organization is as follows: section 2 gives some recent existing approaches using CNNs with LSTM that can learn both ways to predict heart disease. The proposed technique is discussed in section 3. Section 4 analyse the results and analysis of the proposed model. Section 5 concludes the work with some future considerations.

2. LITERATURE REVIEW

Das *et al.* [12] used clinical time-series data to show a hybrid architecture that combines BiLSTM, CNN, and attention-based processes to predict early sepsis. With the help of attention mechanisms, the model can work in on important aspects in the time-series data, leading to better results. Gold and Lawrence [13] described the red deer (RD) algorithm that improves an LSTM model to propose an internet of things (IoT)-enabled healthcare system for identifying cardiac diseases. This data was gathered using wearable health trackers and heart rate monitors. Chichani and Kasar [14] developed an IoT-based HDP model by improving a BiLSTM classifier using Finch hunt optimization. Cloud computing is also employed to handle huge data collected by IoT sensors. This paper demonstrates how the model may use advanced IoT data collection as well as cloud computing to conduct large-scale analyses. The model targets the intricacy of heart disease diagnosis in an effort to enhance forecast accuracy. Lilhore *et al.* [15] addressed the challenge of heart sound classification, particularly in unbalanced datasets. This work proposed a hybrid CNN-Bi-LSTM model that integrates an attention mechanism and synthetic minority over-sampling technique (SMOTE). This study improves classification performance by using SMOTE to fix the class imbalance issue. The CNN takes sequential data, while the Bi-LSTM extracts global and local characteristics. Attention processes enhance the model's accuracy by focusing on the most significant regions of heart sounds [16].

Tran *et al.* [17] implemented CbiLSuccSite algorithm that combines CNN and Bi-LSTM networks and can be used to predict protein succinylation sites from protein sequences. This method encodes protein sequences

using word embedding. Vandermeer and Mahadevan [18] defined the goal of this disease prediction model which is to use patient data to identify potential risks of disease in the future by analyzing time-series data from EHR using LSTM. This study demonstrates that the LSTM is a good fit for EHR time-series data due to its capacity to simulate long-term interdependence. The LSTM outperforms three popular ML models, namely support vector machine (SVM), multi-layer perceptron (MLP), as well as random forest (RF), in disease prediction. Cenitta *et al.* [19] discussed about HDP by a hybrid residual attention-enhanced LSTM (HRAE-LSTM) model. Combining attention processes with residual learning, this model improves LSTM's effectiveness in predicting heart disease. Table 1 gives the advantages and limitations of the existing works.

Table 1. Existing review

Authors	Methodology	Advantages	Limitations
Das *et al.* [12]	Attention-based BiLSTM-CNN for early sepsis prediction	Integrates CNN, BiLSTM, and attention for time-series analysis	High processing requirement and limited generalizability testing
Gold and Lawrence [13]	RD-optimized LSTM integrated with IoT sensors	Supports continuous monitoring and improves diagnosis accuracy	Depends on real-time sensor data and complex data integration
Chichani and Kasar [14]	FHO-modified BiLSTM classifier for HDP	Handles large datasets using cloud computing and provides precise prediction	Requires large cloud infrastructure and handling extensive sensor data
Lilhore *et al.* [15]	CNN-BiLSTM with SMOTE and attention for heart sound classification	Handles class imbalance and captures local and global features	Sensitive to noisy heart sounds and has architectural complexity
Tran *et al.* [17]	CbiLSuccSite using CNN and BiLSTM	Achieved area under the curve (AUC) of 0.826 in 10-fold cross-validation	Domain-specific and difficult to apply to other problems
Vandermeer and Mahadevan [18]	LSTM-based disease prediction using EHR time-series data	Models long-term relationships and outperforms SVM, RF, and MLP	Sensitive to noisy EHR data and requires high-quality labeled data
Cenitta *et al.* [19]	HRAE-LSTM for HDP	Combines attention, residual learning, and LSTM for improved prediction	Requires extensive tuning and considerable computational resources

3. PROPOSED METHOD

Using two open source datasets, this research classifies cardiac disease using a DL-based technique. This work proposes the hybrid CNN-BiLSTM model and is compared with some existing approaches under the same environment. Data preprocessing includes feature selection, normalisation, and SMOTE-based treatment of class imbalances. The CNN extracts spatial characteristics, while LSTM deal with temporal connections. The hybrid CNN-BiLSTM model extracts features by looking at past and future contextual information. The hybrid CNN-BiLSTM model receives particular training optimisations to increase performance and stability, such as learning rate scheduling (ReduceLROnPlateau). The performance measures including F1 score, precision, recall, specificity, as well as accuracy are employed to assess the models. Figure 1 displays the depiction of the proposed hybrid CNN-BiLSTM model.

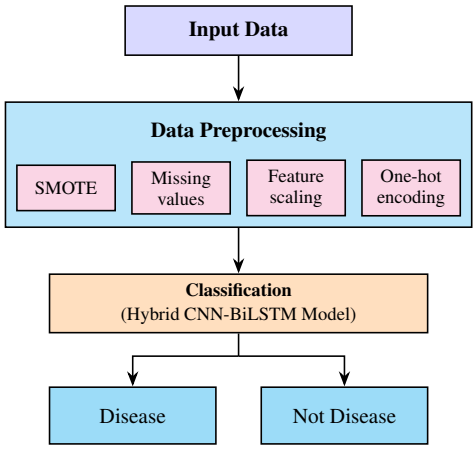


Figure 1. Proposed flow

This study utilizes two publicly available heart disease datasets such as heart failure clinical dataset and heart disease dataset. The Heart Failure Clinical Dataset [20] contains 299 patient records providing 13 clinical features, comprising age, ejection fraction, serum creatinine, diabetes history, and hypertension. It offers helpful information on the chances of death from cardiovascular diseases and mainly focuses on patients who are experiencing heart failure. Many different types of survival rate predictions and risk factor analyses make use of this dataset.

In the Heart Disease Dataset [21] types of chest discomfort, blood pressure, cholesterol, and fasting blood sugar are some of the 14 characteristics included in this dataset that pertain to 1025 occurrences. This dataset is excellent for binary classification tasks (disease vs. no disease) as it offers a more comprehensive classification of heart disease existence than the heart failure dataset. It is useful for determining which variables are most likely to cause cardiovascular problems. Both datasets undergo preprocessing steps to improve data quality.

3.1. Data preprocessing

Initially, the datasets were examined for missing and inconsistent values. The non-numeric values, were converted into numerical format, and missing entries were removed to maintain data integrity. Feature standardization was then done using the StandardScaler method to normalize feature distributions and improve convergence during network training. In this work, outlier handling was not performed explicitly, as the selected benchmark datasets contain clinically validated records with limited extreme observations, and feature normalization minimizes the influence of value magnitude differences.

Alternatively, one-hot encoding is employed to numerically signify categorical qualities in categorical variables. Finally, synthetic minority over-sampling technique (SMOTE) checks class imbalance during cross-validation. The datasets may have class imbalance which can cause the proposed model to make biased predictions. The SMOTE technique makes use of minority class synthetic samples to address this problem. The default configuration with $k = 5$ nearest neighbors were used to produce synthetic minority-class samples. The oversampling ratio was determined automatically to attain balanced class distributions without altering the original testing data. Let $\mathbf{x}_i$ be a minority class sample, and $\mathbf{x}_{nn}$ be a k-nearest neighbor. A new synthetic sample $\mathbf{x}_{new}$ is given by (1):

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda \left(\mathbf{x}_{nn} - \mathbf{x}_i \right), \tag{1}$$

where λ is a random number in the range $[0, 1]$, confirming the new sample is a convex combination of existing samples. This technique effectively balances the dataset and improves model generalization.

3.2. Model architectures

This research tests various DL models for the classification of heart diseases.

3.2.1. Convolutional neural network

CNNs extract spatial features from organized medical data. A single 1D convolutional layer is used to apply filters to the input feature vector and find patterns. Let $\mathbf{x}$ be the input feature vector, $\mathbf{w}$ is the filter, and b be the bias. The convolution process is given by (2):

$$y_i = f\left(\sum_{j=0}^{K-1} w_j x_{i+j} + b \right), \tag{2}$$

where $f(\cdot)$ denotes the activation function (ReLU). MaxPooling is applied to decrease dimensionality, followed by dense layers with 16 filters. The CNN fails to deal with sequential dependencies.

3.2.2. Long short-term memory

LSTMs capture temporal dependencies in sequential data. Each LSTM unit has three gates as in (3) to (5). Forget gate controls what information to discard.

$$f_t = \sigma \left(W_f [h_{t-1}, x_t] + b_f \right), \tag{3}$$

Input gate determines what new information to store in (4):

$$i_t = \sigma \left(W_i [h_{t-1}, x_t] + b_i \right), \tag{4}$$

Output gate produces the final hidden state in (5):

$$o_t = \sigma\left(W_o[h_{t-1}, x_t] + b_o\right), \tag{5}$$

where h_t gives the hidden state at time t and c_t is the cell state. The LSTM drawback is it only process in single direction.

3.2.3. Convolutional neural network long short-term memory

In CNN-LSTM, sequential learning (LSTM) and spatial feature extraction (CNN) are combined. After LSTM checks the incoming data for sequential dependencies, the CNN processes it to extract local patterns. The key difference is that LSTM operates on CNN feature maps $\mathbf{F}_t$ instead of raw input $\mathbf{x}_t$ in (6):

$$h_t = \mathrm{LSTM}\left(\mathbf{F}_t, h_{t-1}\right), \tag{6}$$

where $\mathbf{F}_t$ is the feature representation at time step t. CNN-LSTM lacks bidirectional processing. The CNN-LSTM model combines convolutional feature extraction with a 32-unit single LSTM layer.

3.2.4. Proposed hybrid convolutional neural network–bidirectional long short-term memory model

The proposed hybrid CNN-BiLSTM model combines CNN with BiLSTM considering spatial feature extraction and temporal sequence modeling, taking use of past and future dependencies for better accuracy [22]–[25]. Input layer accepts structured numerical features and CNN block comprises Conv1D layer and MaxPooling layer. The two Conv1D layer used with 64 and 128 filters extracts spatial patterns using convolution filters. Further batch normalization and max pooling layers follows it to advance training stability, feature extraction, and reduce dimensionality. The BiLSTM block has forward LSTM, backward LSTM and concatenation layer. The forward LSTM processes sequence in forward direction while the reverse direction is in the backward LSTM. This BiLSTM layer with 64 units captures temporal dependencies in both forward and backward directions.

Finally, the concatenation layer merges bidirectional outputs. The overfitting issue is sorted by using the dropout layers after BiLSTM layer with drop-rate of 0.4. The output dense layer comprises of a sigmoid activation function since the two datasets support binary classification. The feature extraction takes place and passes through a dense layer with 64 neurons and ReLU activation. Finally, binary classification is done by the output layer which uses a sigmoid activation function. Let $\overrightarrow{h_t}$ as well as $\overleftarrow{h_t}$ be the forward and backward hidden states at time t in (7) and (8):

$$\overrightarrow{h_t} = \mathrm{LSTM}_f\left(F_t, \overrightarrow{h}_{t-1}\right), \tag{7}$$

$$\overleftarrow{h_t} = \mathrm{LSTM}_b\left(F_t, \overleftarrow{h}_{t+1}\right). \tag{8}$$

The final hidden state representation is in (9):

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right], \tag{9}$$

where h_t is passed to fully connected layers for classification. The proposed model does better classification as it uses bidirectional processing, which picks up richer sequential patterns.

3.3. Training and evaluation

The models training is done using 5-fold stratified cross-validation which confirms robustness. Along with learning rate scheduler, ReduceLROnPlateau, the proposed CNN-BiLSTM model employs the Adam optimizer. Table 2 displays the hyperparameter configurations for proposed and existing models. For the proposed model, the 0.0005 learning rate was selected after preliminary experiments as it provided a stable convergence rate while avoiding oscillations during optimization. Dropout was configured to balance the model's complexity and generalization performance. Higher dropout rates led to underfitting, while lower values increased the risk of overfitting. The values used achieved stable validation performance. Although a sensitivity analysis was not considered, preliminary experiments with different learning rates and dropout values showed that the selected hyperparameter configuration consistently provided stable convergence and competitive classification performance across both benchmark datasets.

The cross-validation results average and is reported as mean ± standard deviation to reproduce model stability across different folds. A paired t-test was also done which evaluates the proposed model statistical significance in comparison to other baseline models.

Table 2. Hyperparameter configuration

Parameter	CNN	LSTM	CNN-LSTM	Proposed CNN-BiLSTM
Batch size	32	64	64	64
Learning rate	0.001	0.001	0.0005	0.0005
Optimizer	Adam	Adam	Adam	Adam
Dropout rate	0.3	0.4	0.5	0.5
Epochs	50	50	50	50

4. RESULTS

The experimentation is done with an NVIDIA GPU, 32 GB of RAM, as well as an Intel Core i5 CPU. Then for data preprocessing and further training the Python libraries TensorFlow/Keras, Scikit-learn, and Pandas are used. The average training time was approximately 2–4 seconds per epoch.

4.1. Evaluation metrics

To estimate the proposed and the existing model's performance, the subsequent metrics were considered. Accuracy estimates overall predictions perfection as in (10):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP is the true positive, TN denotes true negative, FP signify false positive, and FN denote false negative. Precision represents the amount of properly predicted positive cases out of total predicted positive cases as in (11):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

Recall estimates the ability of the model to classify all relevant cases as in (12):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

F1-score is a harmonic mean of precision with recall, matching both metrics as in (13):

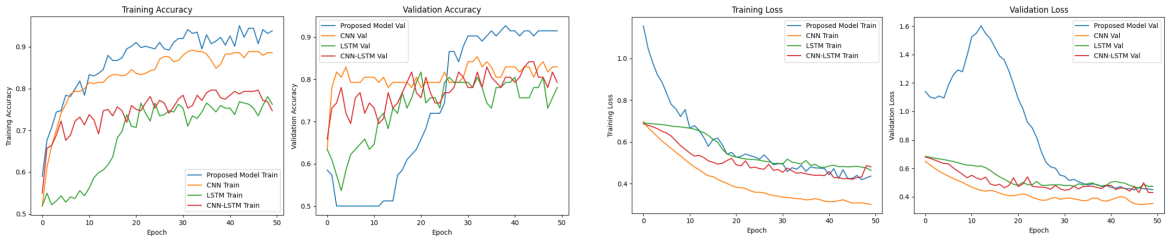
$$F1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

Specificity gives the proportion of actual negatives correctly predicted as in (14):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{14}$$

4.2. Performance comparison

The proposed hybrid CNN-BiLSTM model does better for dataset 1 in terms of validation accuracy (98.3%), and validation loss. The F1-score of 98.5% demonstrates a balanced precision-recall tradeoff. The findings of dataset 2 also demonstrate that BiLSTM is superior to other models, with a 98.6% accuracy rate. Important for medical uses, the 99% recall guarantees few FNs. Figure 2 shows the training along with validation accuracy/loss curves for dataset 1. In dataset 1 as in Figure 2(a), CNN and LSTM models reach lower plateaus, but proposed hybrid CNN-BiLSTM model consistently shows training accuracy, reaching 98.3%. Consistent with this trend, the validation accuracy verifies improved generalizability. With a training loss of 0.02 epochs as in Figure 2(b), proposed hybrid CNN-BiLSTM is clearly an efficient learner. The CNN's and LSTM's changing loss values demonstrate unstable optimization. As a result of its lower validation loss of 0.05 compared to other models' losses of 0.1–0.15, proposed hybrid CNN-BiLSTM is more resistant to overfitting.



(a) (b)

Figure 2. Training and validation performance curves for dataset 1: (a) accuracy curves and (b) loss curves

The 5-fold cross-validation comparison of the proposed model is given in Table 3 on dataset 1. The table confirms that the proposed model achieved the high accuracy of 0.8194 ± 0.0213, outperforming other existing models. The proposed model's F1-score, demonstrates its superior performance. Notably, higher precision of the proposed model suggests its improved reliability in positive predictions. While CNN showed slightly higher recall, it has lower precision and higher variance, showing less stable performance. Furthermore, the high specificity demonstrates its effectiveness in correctly identifying negative cases.

Table 3. 5-fold cross validation results dataset 1

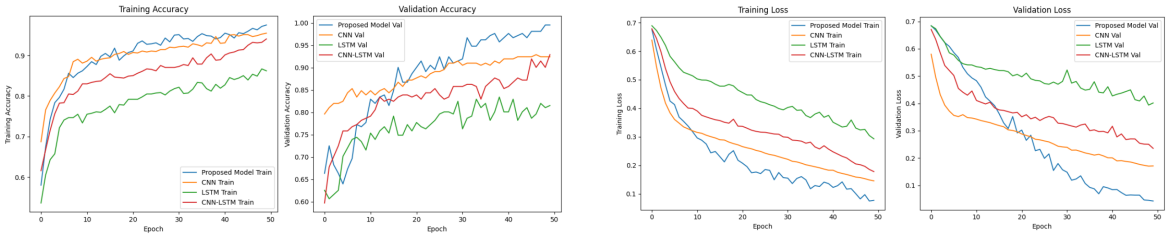
Models	Accuracy	Precision	Recall	F1-score	Specificity
Proposed	0.8194 ± 0.0213	0.7150 ± 0.0308	0.7289 ± 0.0452	0.7215 ± 0.0328	0.8620 ± 0.0230
CNN	0.7527 ± 0.0761	0.6054 ± 0.1208	0.8226 ± 0.0608	0.6876 ± 0.0541	0.7194 ± 0.1399
LSTM	0.6825 ± 0.1020	0.5220 ± 0.1013	0.7500 ± 0.0778	0.6087 ± 0.0745	0.6510 ± 0.1528
CNN-LSTM	0.7825 ± 0.0493	0.6314 ± 0.0656	0.7805 ± 0.0705	0.6976 ± 0.0661	0.7833 ± 0.0466

Table 4 presents the paired t-tests results from 5-fold cross-validation for dataset 1. This statistical analysis confirms that the proposed model attains a statistically significant enhancement over the LSTM model. However, the differences between the proposed model with CNN and CNN-LSTM are not statistically significant. This proposes that while the proposed model outperforms these models by mean performance, it is not constant enough across folds to reach statistical significance. These results thus show the competitive and stable performance of the proposed CNN–BiLSTM model.

Table 4. Paired t-tests dataset 1

Models	t-statistic	p-value	Significance
Proposed vs. CNN	2.4852	0.067830	Not statistically significant ($p \geq 0.05$)
Proposed vs. LSTM	2.9771	0.040857	Statistically significant difference ($p < 0.05$)
Proposed vs. CNN-LSTM	1.6238	0.179743	Not statistically significant ($p \geq 0.05$)

Figure 3 shows the training along with validation accuracy/loss curves for dataset 2. When tested on dataset 2, as in Figure 3(a) proposed model outperforms both CNN (95.2%) and LSTM (96.3%), with 98.6% accuracy. In contrast to CNN and LSTM as in Figure 3(b), where loss values sometimes spike, proposed hybrid CNN-BiLSTM displays a decreasing trend, stabilizing around 0.03. Reinforcing its steady convergence, the validation loss is still less than 0.06. These findings demonstrate that proposed hybrid CNN-BiLSTM can consistently generalize well across a wide variety of datasets. Stability is shown by the smooth convergence and minimal fluctuations in the proposed hybrid CNN-BiLSTM model. Table 5 presents the performance comparison between proposed and baseline models on dataset 2. In comparison to all three existing models, the proposed model reached the highest value for all metrics, showing excellent classification performance. This improvement confirms the robustness and the low standard deviation values across these metrics points the stability and reliability of the proposed model.



(a) (b)

Figure 3. Training and validation performance curves for dataset 2: (a) training and validation accuracy curves and (b) training and validation loss curves

Table 5. 5-fold cross validation results dataset 2

Models	Accuracy	Precision	Recall	F1-score	Specificity
Proposed	0.9854 ± 0.0142	0.9867 ± 0.0145	0.9848 ± 0.0159	0.9857 ± 0.0139	0.9860 ± 0.0152
CNN	0.9112 ± 0.0167	0.9102 ± 0.0253	0.9182 ± 0.0209	0.9139 ± 0.0158	0.9038 ± 0.0288
LSTM	0.8049 ± 0.0639	0.8260 ± 0.0750	0.7889 ± 0.0539	0.8064 ± 0.0597	0.8217 ± 0.0857
CNN-LSTM	0.9171 ± 0.0324	0.9191 ± 0.0363	0.9201 ± 0.0396	0.9192 ± 0.0322	0.9138 ± 0.0405

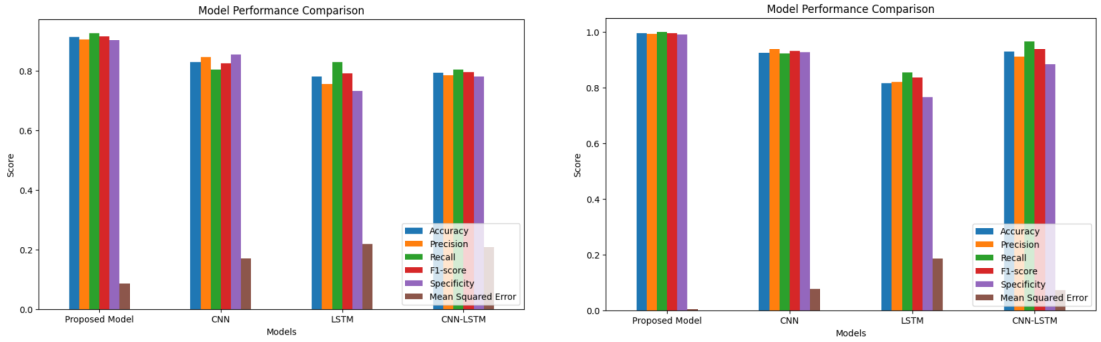
The statistical significance evaluation for dataset 2 is reported in Table 6. Paired comparisons obtained from the 5-fold cross-validation results show that the proposed CNN–BiLSTM differs significantly from CNN, LSTM, and CNN-LSTM, with all reported p-values below 0.05. This statistical evidence supports the performance differences observed between the proposed model and the three baseline models on dataset 2.

Table 6. Paired T-tests dataset 2

Models	t-statistic	p-value	Significance
Proposed vs. CNN	6.8990	0.002315	Statistically significant difference ($p < 0.05$)
Proposed vs. LSTM	4.6278	0.009823	Statistically significant difference ($p < 0.05$)
Proposed vs. CNN-LSTM	4.6108	0.009950	Statistically significant difference ($p < 0.05$)

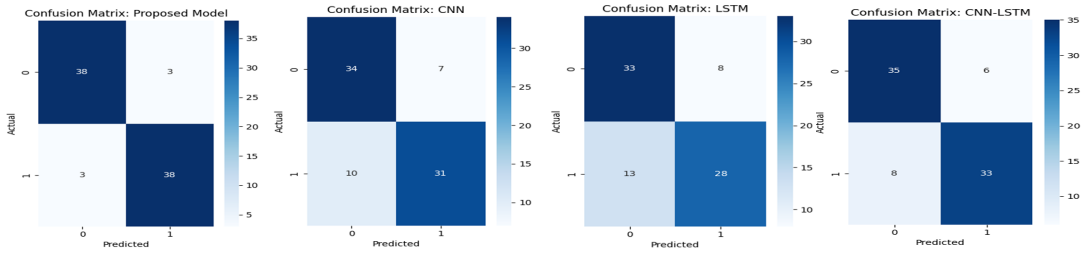
Figure 4 summarizes the comparative classification results obtained from datasets 1 and 2 and is intended to show the relative behaviour of the evaluated models across the main performance metrics. Figure 4(a) presents the results for dataset 1, where the proposed CNN–BiLSTM records an accuracy of 98.1%, recall of 98.0%, and F1-score of 98.0%. CNN and LSTM produce lower recall values of approximately 94–95%, indicating that a larger proportion of heart disease cases are missed by these models. The specificity of the proposed model reaches 97.8%, compared with 95.6% for LSTM and 94.2% for CNN, which corresponds to fewer FP predictions. Figure 4(b) reports the comparison for dataset 2. The proposed CNN–BiLSTM attains an F1-score of 98.4%, compared with 96.2% for CNN and 97.1% for LSTM, while its recall of 99.0% indicates that only a small number of positive cases are misclassified.

Figure 5 complements these metric-based comparisons by presenting the confusion matrices for dataset 1, allowing the numbers of TP, TN, FP, and FN predictions to be examined separately for each evaluated model. Figure 5(a) presents the confusion matrix of the first baseline model and shows the distribution of its correct and incorrect classifications. Figure 5(b) presents the corresponding confusion matrix for the second baseline model, providing a separate view of its positive and negative prediction errors. Figure 5(c) reports the confusion matrix for the third baseline model and shows its class-wise prediction distribution. Figure 5(d) presents the confusion matrix of the proposed CNN–BiLSTM model, where the lower FP rate is consistent with its ability to distinguish heart disease cases from non-heart-disease cases. The corresponding accuracy of 98.3% indicates that most samples in dataset 1 are classified correctly.



(a) (b)

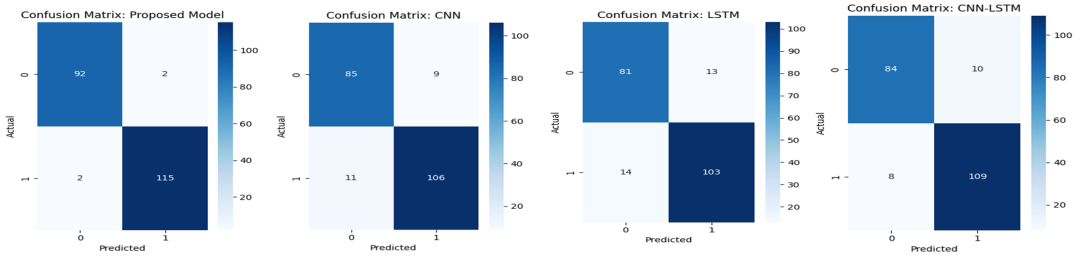
Figure 4. Comparison graphs: (a) dataset 1 and (b) dataset 2



(a) (b) (c) (d)

Figure 5. Confusion matrix for dataset 1: (a) proposed model, (b) CNN, (c) LSTM, and (d) CNN-LSTM

Figure 6 presents the confusion matrices for dataset 2 and is intended to show the distribution of TP, TN, FP, and FN predictions for the proposed CNN–BiLSTM and the baseline models. Figure 6(a) presents the confusion matrix of the first baseline model and shows the number of correctly and incorrectly classified samples. Figure 6(b) reports the corresponding prediction distribution for the second baseline model, allowing its FP and FN errors to be examined separately. Figure 6(c) presents the confusion matrix of the third baseline model and provides another class-level comparison against the proposed approach. Figure 6(d) shows the confusion matrix of the proposed CNN–BiLSTM model, which attains a TP rate of 98.9%, compared with 96.5% for CNN and 97.3% for LSTM. The reduction in FN predictions is consistent with this higher TP rate, while a specificity of 98.6% indicates that most non-disease cases are also classified correctly.

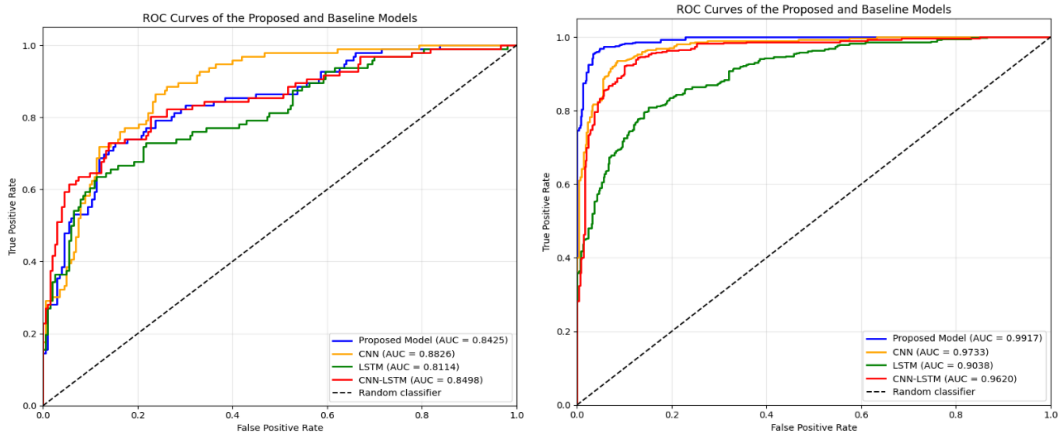


(a) (b) (c) (d)

Figure 6. Confusion matrix for dataset 2: (a) proposed model, (b) CNN, (c) LSTM, and (d) CNN-LSTM

Figure 7 presents the ROC analysis and compares the discriminative behaviour of the proposed model with the baseline models across different decision thresholds. Figure 7(a) illustrates the ROC curves of the proposed model and the baseline models on dataset 1. The proposed CNN–BiLSTM model attained an AUC of

0.8425, representing good insight between patients with and without heart failure. Although the CNN model obtained the highest AUC, its overall classification performance was lower, signifying that a higher AUC alone does not translate into balanced classification performance. These results prove good discriminative ability of the proposed model while preserving a balanced trade-off among accuracy, precision, recall, and specificity across cross-validation folds. Figure 7(b) presents the ROC curves obtained for dataset 2.



(a) (b)

Figure 7. ROC curve: (a) dataset 1 and (b) dataset 2

The proposed model achieved the highest AUC value of 0.9917, demonstrating good discriminative capability for distinguishing heart disease cases from healthy individuals. The CNN and CNN–LSTM models had strong performance, with AUC values of 0.9733 and 0.9620, whereas the LSTM model got the lowest AUC (0.9038). The proposed model ROC curve consistently remains closer to the upper-left corner, demonstrating higher sensitivity while maintaining a lower FP rate over a wide range of decision thresholds. The ROC analysis further demonstrates the proposed CNN–BiLSTM model robustness across different classification thresholds. From a clinical perspective, keeping a high TP rate (recall) is important because FN predictions may delay the analysis and treatment of patients with heart disease. Also, the proposed model achieves competitive specificity, thereby reducing needless follow-up examinations caused by FP predictions. This balanced trade-off between sensitivity and specificity makes the proposed approach appropriate for supporting clinical decision-making in heart disease diagnosis.

4.3. Discussion

Due to bidirectional feature extraction, the proposed hybrid CNN-BiLSTM model outperforms the other existing models. An important component of medical classification tasks, this bidirectional learning greatly decreases FPs and FNs. During extraction of temporal links, the hybrid CNN-BiLSTM model outperforms CNN and CNN-LSTM because of its superior handling of sequential dependencies. Validation loss reductions provide evidence of its resilience. It was found that recall and specificity are better, which means that fewer mistakes will be made when heart diseases are being found. The results show that hybrid CNN-BiLSTM can handle both classic classification tasks and more complex real-time monitoring problems. Its versatility across datasets further enhances the model's use in many healthcare contexts. The hybrid CNN-BiLSTM is able to improve its prediction capabilities by collecting long-range relationships via its bidirectional feature extraction. The comparative examination confirms its appropriateness for medical applications, demonstrating continuous improvements. This proposed model has the ability to be incorporated into healthcare monitoring systems and aid physicians in heart diseases diagnosis.

The proposed CNN–BiLSTM model high recall can be attributed to the bidirectional processing capability of the BiLSTM layer. The bidirectional learning allows to capture rich feature dependencies, thereby improving its capability to recognize positive heart disease cases. Thus, the proposed model reduces FN predictions and achieves higher recall compared with the existing models. Although the proposed CNN–BiLSTM model presents additional computational complexity compared with CNN or LSTM models, this

increase is justified by the substantial improvement in classification performance. The bidirectional recurrent layer needs additional memory and processing time during training. Still, the resulting gains in performance metrics demonstrate that the additional computational overhead is satisfactory for clinical decision-support systems where diagnostic reliability is prioritized over training efficiency. From a clinical view, recall is significant since FN predictions may delay diagnosis and treatment for patients with heart disease. Although specificity is important to reduce redundant follow-up examinations, prioritizing higher recall improves patient safety by minimizing missed disease cases. Therefore, the proposed CNN–BiLSTM model gives balanced trade-off between sensitivity and specificity, making it appropriate for decision-support applications. Apart from heart disease diagnosis, the proposed CNN–BiLSTM framework can be extended to wearable healthcare devices including IoT-enabled patient monitoring systems, and EHR time-series analysis, where accurate sequential pattern learning is important. This study is limited to only two publicly available Kaggle datasets and does not consider external clinical validation.

5. CONCLUSION

This research introduces a hybrid CNN-BiLSTM model which efficiently captures sequential and spatial patterns. The convolutional layers capture local feature relationships, while the BiLSTM layer models dependencies in forward and backward directions. Experimental evaluation was done on two benchmark heart disease datasets. The proposed model achieved 98.54% accuracy, 98.48% recall, and 98.57% F1-score on the heart disease dataset. On the heart failure clinical dataset, the model achieved an 81.94% average cross-validation accuracy, demonstrating consistent and reliable performance. The results demonstrate that hybrid architectures are beneficial for complex medical datasets, also performance varies depending on dataset characteristics and size. The comparison done show that hybrid CNN-BiLSTM is good than CNN, LSTM, and CNN-LSTM models. The comparisons in the bar graph demonstrate its excellent precision-recall balance, which adds credence to its possible use in clinical settings. Statistical significance was done using paired t-tests which confirm statistically significant improvements for both the datasets. From a clinical perspective, the proposed model provides improved recall, which reduces the likelihood of FN predictions and supports earlier identification of patients at risk. Integration with wearable devices and real-time deployment will be done in future improvements. Furthermore, the future work will validate real-world hospital datasets and combining attention mechanisms to improve model interpretability. Also, the framework will be extended for edge-device deployment and multimodal healthcare applications.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohammad Azhar	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Lazer Mary Gladence		✓				✓		✓	✓	✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY STATEMENT

This work uses data from https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data [19] and https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset [20].

REFERENCES

[1] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "A hybrid bidirectional LSTM and 1D CNN for heart disease prediction," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 21, no. 10, pp. 135–142, Oct. 2021, doi: 10.1109/iceem52022.2021.9480625.

[2] S. U. Hassan, M. S. M. Zahid, T. A. A. Abdullah, and K. Husain, "Classification of cardiac arrhythmia using a convolutional neural network and bi-directional long short-term memory," *Digital Health*, vol. 8, pp. 1–10, 2022, doi: 10.1177/20552076221102766.

[3] P. Dileep *et al.*, "An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7253–7266, 2023, doi: 10.1007/s00521-022-07064-0.

[4] M. Alkhodari and L. Fraiwan, "Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings," *Computer Methods and Programs in Biomedicine*, vol. 200, Jul. 2021, doi: 10.1016/j.cmpb.2021.105940.

[5] P. K. Shrivastava, M. Sharma, and A. Kumar, "HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithms," *Measurement: Sensors*, vol. 25, pp. 100657, 2023, doi: 10.1016/j.measen.2022.100657.

[6] U. K. Lilhore *et al.*, "A deep learning approach for heart disease detection using a modified multiclass attention mechanism with BiLSTM," *Scientific Reports*, vol. 15, no. 1, pp. 25273, 2025, doi: 10.1038/s41598-025-09594-8.

[7] M. Najia and F. Benzarti, "An enhanced hybrid model combining CNN, BiLSTM, and attention mechanism for ECG segment classification," *Biomedical Engineering and Computational Biology*, vol. 16, 2025, doi: 10.1177/11795972251341051.

[8] S. Hiriyannaiah, S. GM, K. MHM, and K. G. Srinivasa, "A comparative study and analysis of LSTM deep neural networks for heartbeats classification," *Health and Technology*, vol. 11, no. 3, pp. 663–671, 2021, doi: 10.1007/s12553-021-00552-8.

[9] P. Dhaka and R. Sehrawat, "Adaptive ensembled fusion based deep CNN-BiLSTM model for heart disease prediction in IoT," *Fusion: Practice and Applications*, vol. 14, no. 1, pp. 1–10, 2024, doi: 10.54216/fpa.140104.

[10] C. Zhou *et al.*, "A comprehensive review of deep learning-based models for heart disease prediction," *Artificial Intelligence Review*, vol. 57, no. 10, 2024, doi: 10.1007/s10462-024-10899-9.

[11] P. V. Mulwani, M. Bhende, and S. Sharma, "Revolutionizing heart disease care with AI: Precision medicine in diagnosis, prediction, and treatment," in *Responsible AI for Digital Health and Medical Analytics*, Hershey, PA, USA: IGI Global, 2025, pp. 465–494, doi: 10.4018/979-8-3693-6294-5.ch019.

[12] P. P. Das *et al.*, "An attention-based bidirectional LSTM-CNN architecture for the early prediction of sepsis," *International Journal of Data Science and Analytics*, vol. 20, no. 3, pp. 1–15, 2024, doi: 10.1007/s41060-024-00568-z.

[13] O. C. Gold and J. Lawrence, "Enhanced LSTM for heart disease prediction in IoT-enabled smart healthcare systems," *The Scientific Temper*, vol. 15, no. 2, pp. 2238–2247, doi: 10.58414/scientifictemper.2024.15.2.37.

[14] Y. S. Chichani and S. L. Kasar, "An efficient IoT enabled heart disease prediction model using Finch hunt optimization modified BiLSTM classifier," *Biomedical Signal Processing and Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.107170.

[15] U. K. Lilhore, S. Simaiya, M. Alhussein, S. Dalal, K. Aurangzeb, and A. Hussain, "An attention-driven hybrid deep neural network for enhanced heart disease classification," *Expert Systems*, vol. 42, no. 2, 2025, doi: 10.1111/exsy.13791.

[16] B. Xia, N. Innab, V. Kandasamy, A. Ahmadian, and M. Ferrara, "Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization," *Scientific Reports*, vol. 14, no. 1, pp. 21777, 2024, doi: 10.1038/s41598-024-71932-z.

[17] T.-X. Tran, N. Q. K. Le, and V.-N. Nguyen, "Integrating CNN and Bi-LSTM for protein succinylation sites prediction based on natural language processing technique," *Computers in Biology and Medicine*, vol. 186, 2025, doi: 10.1016/j.compbiomed.2025.109664.

[18] E. Vandermeer and K. Mahadevan, "Disease prediction using electronic health records with long short-term memory networks," *Journal of Computer Technology and Software* , vol. 4, no. 1, pp. 1–8, 2025, doi: 0.5281/zenodo.14785344.

[19] D. Cenitta, R. V. Arjunan, G. Paramasivam, N. Arul, A. Palkar, and K. Chadaga, "Ischemic heart disease prognosis: A hybrid residual attention-enhanced LSTM model," *IEEE Access*, vol. 13, pp. 4281-4289, 2025, doi: 10.1109/ACCESS.2024.3524604.

[20] A. Mvd, "Heart failure clinical data," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data. (Accessed: Sept. 24, 2025).

[21] J. Smith, "Heart disease dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. (Accessed: Sept. 24, 2025).

[22] E. Kalatehjari, M. M. Hosseini, A. Harimi, and V. Abolghasemi, "Advanced ensemble learning-based CNN-BiLSTM network for cardiovascular disease classification using ECG and PCG signal," *Biomedical Signal Processing and Control*, vol. 108, pp. 107846, 2025, doi: 10.1016/j.bspc.2025.10784.

[23] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Computers in Biology and Medicine*, vol. 146, pp. 105624, 2022, doi: 10.1016/j.compbiomed.2022.105624.

[24] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation," *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 31759–31773, 2023, doi: 10.1007/s11042-023-14817-z.

[25] V. Chole, M. Thawakar, M. Choudhari, S. Chahande, S. Verma, and A. Pimpalkar, "Enhancing heart disease risk prediction with GdHO fused layered BiLSTM and HRV features: A dynamic approach," *Biomedical Signal Processing and Control*, vol. 95, pp. 106470, 2024, doi: 10.1016/j.bspc.2024.106470.

BIOGRAPHIES OF AUTHORS



Mohammad Azhar is a researcher at the Sathyabama Institute of Science and Technology in Chennai and holds an M.E. degree from the Sathyabama Institute of Science and Technology in Chennai. He is very interested in machine learning and deep learning, much more on the technical side. He can be contacted at email: azharmdcse@gmail.com.



Lazer Mary Gladence is an associate professor at the School of Computing at the Satyabama Institute of Science and Technology in Chennai, India. She defended her doctoral thesis. She received her Ph.D. in Data Mining from the Satyabama Institute of Science and Technology in 2017. Her research interests include deep learning, artificial intelligence, data mining, machine learning, biocomputing, and data analytics. She has over 70 publications in these fields. Unnat Bharat Abhiyan (UBA), MHRD, Government of India. She can be contacted at email: lgladence@gmail.com.