International Journal of Health Sciences and Research
[Indexed, Peer-Reviewed Journal]
Volume 16; Issue: 5; May 2026
Website: www.ijhsr.org
Original Research Article *ISSN: 2249-9571*

Systematic Benchmarking of ANN, CNN, and LSTM Models for Heart Disease Prediction Using Multimodal Clinical Features

Sunanda Budhal[1,2], Sheetalrani Kawale[2], Bhagirathi Hallalli[1], Nitin Agarwal[3]

[1]Department of Computer Science, Government First Grade College, Bagalkot-587103, Karnataka, India
[2]Department of Computer Science, Karnatak State Women's University, Vijayapura-596101, Karnataka, India
[3]Consulting Physician and Cardiologist, Ayush Multi Speciality Hospital and Research Centre (AMSHRC) Pvt. Ltd., Vijayapura, Karnataka, India

Corresponding Author: Sunanda Budhal

DOI: https://doi.org/10.52403/ijhsr.20260521

ABSTRACT

Accuracy and early prediction of heart disease are essential for efficient clinical decision-making and risk management. This paper presents a deep learning-based predictive approach for heart disease diagnosis using a real-time clinical dataset Ayush Multi Speciality Hospital and Research Centre (AMSHRC) Pvt. Ltd. Vijayapura, Karnataka, India augmented with sophisticated diagnostic attributes. The proposed approach combines systematic data preprocessing, feature extraction, and three different neural network architectures: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) networks, to compare the prediction accuracy and robustness. Five-fold cross-validation stratified and independent test-set validation are performed. Cross-validation results show that CNN yields the highest validation accuracy (96.19% +0.61), MCC (0.9087 +0.0144) and ANN yields equal validation accuracy (95.87% +0.82%), highest recall value (94.92% +1.36%) and highest AUC (0.9915 +0.0020). The performance on test-sets also verifies that ANN is the most appropriate model overall and has an accuracy of 96.20 and AUC of 0.9936. The inclusion of ECG, ECHO, TMT, and CAG features is a very good approach and helps in improving the prediction accuracy, and it also verifies that ANN is a good model that can be used in practice.

Keywords: Cardiovascular disease (CVDS), long short-term memory (lSTM), deep learning (DL), convolutional neural network (CNN), artificial neural network (ANN), heart disease prediction

INTRODUCTION

The World Health Organization says that the cardiovascular diseases (CVDs) are still the leading cause of death in the world. CVDs are the account for a significant proportion of deaths globally [1]. Heart disease remains a major public health issue because it is characterized by high prevalence and often asymptomatic progression during its early stages [2]. Therefore, early diagnosis and timely intervention have become one of the ways in reducing mortality rates and improving patient wellbeing.

Traditional diagnostic and statistical risk assessment methods were limited by heavy reliance on manually selected clinical features and expert knowledge, lacking the capability to design a complex nonlinear

relationships model is present in real-world clinical data [3-4]. With the wide availability of electronic health records, machine learning and deep learning techniques have been increasingly adopted for automated heart disease prediction. Recent related studies and surveys report that data-driven approaches often outperform traditional methods when trained on comprehensive clinical features.

CNNs are especially effective at learning hierarchical feature representations, particularly from structured or signal-based data [5]; the LSTM networks can handle dependencies in clinical data [6]. However, most of the previous studies either target classical ML models or assess a single deep learning architecture without allowing fair comparison and generalization. Besides, the performance assessment in previous studies was often limited to only one validation strategy, and the application of clinically relevant metrics, such as the MCC and cross-validated ROC-AUC, still remains limited [7]. Recently, a variety of wearable devices have emerged, increasing cardiac monitoring capacity and thus continuous health assessment [8]. However, clinical decision support with reliable predictive models requires robust training on comprehensive diagnostic data.

In our previous work, Budhal, Sunanda et al. [9] performed heart disease detection using machine learning algorithms on 1,100 patient records. Although promising results were obtained, generalization and representation learning were at a threshold due to the limited dataset and reliance on ML models. Therefore, the present study extends prior work through the adoption of deep learning architectures on an expanded dataset of 2,000 real-time clinical records gathered from AMSHRC Pvt.Ltd.

Various deep learning architectures were also extensively examined in cardiovascular diseases detection problems. The literature exhibits differences in the predictive performance of the models due to the data sizes and the approaches considered. Recent literature highlights the importance of considering sample sizes, models, as well as the considered accuracy [10-20]. However, despite these advances, existing research shows a lack in systematic comparisons between ANN, CNN, and LSTM architectures. Most existing research makes use of standard data sets such as UCI heart disease data repository [16], Kaggle data sets [17], etc., often use simple diagnostic attributes.

This study presents a systematic comparative analysis of ANN, CNN, and LSTM models for early heart disease detection using structured clinical data. A unified preprocessing pipeline incorporating missing value imputation, feature scaling, and categorical encoding is employed. Model performance is evaluated through both stratified hold-out testing and stratified five-fold cross-validation, considering multiple clinically relevant metrics to ensure a balanced and reliable assessment.

The rest of the paper is organized as follows. Section II presents the Materials and Methods while Section III presents the proposed Results, Section IV presents Discussion. Finally, Section V presents conclusion, along with the clinical implications and future work.

MATERIALS & METHODS

This section outlines the proposed deep learning method used to predict early instances of heart disease using a real-time data set containing information of 2000 patients collected at AMSHRC Pvt. Ltd. The proposed methodology works in preprocessing, Training, Testing and Validation, as well as the evaluation through the construction of models based on ANN, CNN and LSTM models. The following Figure 1 shows an overview of the proposed methodology for heart disease prediction using deep learning models.

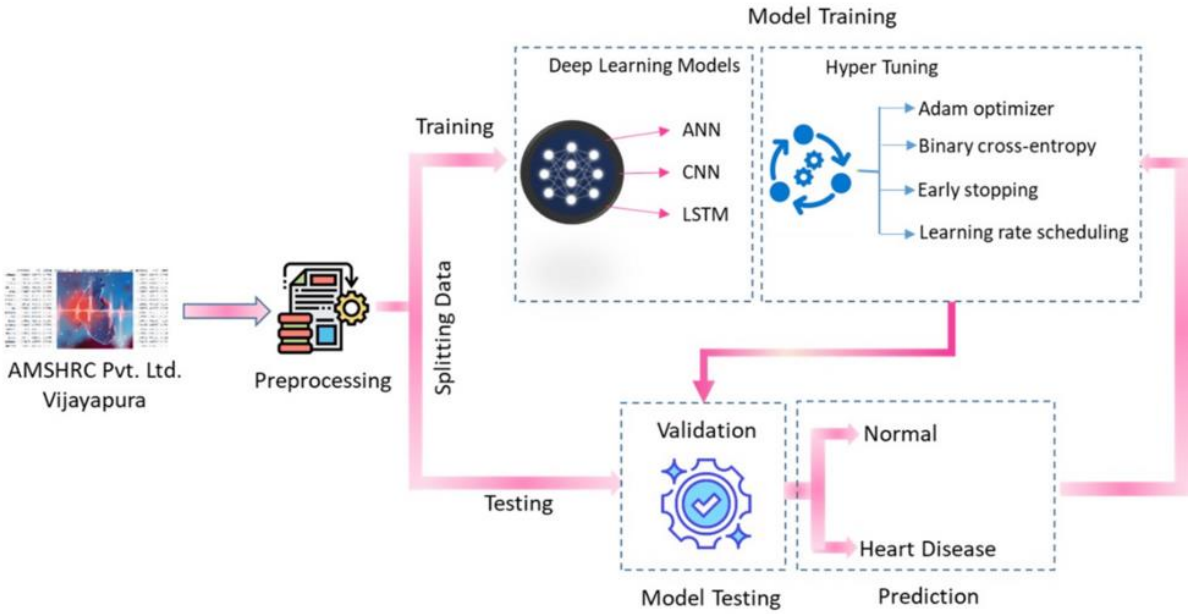

Figure 1: Proposed Methodology

Dataset Description and acquisition

In the proposed research work, clinical dataset of 2000 patient records collected from AMSHRC Pvt.Ltd. The dataset contains demographic information, routine clinical parameters, and advanced cardiological test results in the form of ECG, ECHO, TMT, and CAG reports. Patient privacy is ensured by anonymizing all the records before processing [9].

Data Preprocessing

The collected clinical dataset contained both numerical and categorical attributes, along with missing values, which required careful preprocessing prior to model training. Numerical features, including age, systolic blood pressure, diastolic blood pressure, pulse rate, oxygen saturation, and body weight, were processed separately from categorical attributes such as gender, chest pain type, diabetes status, hypertension, lifestyle habits, and advanced diagnostic test outcomes.

Missing values in numerical features were imputed using the median strategy to reduce the influence of outliers, while missing categorical values were replaced using the most frequent category. All numerical attributes were standardized using z-score normalization to ensure uniform feature scaling and to improve model convergence during training. Categorical variables were transformed using one-hot encoding to convert nominal values into a machine-readable binary format, while preserving category information and avoiding ordinal assumptions.

A unified preprocessing pipeline was developed with the use of a column-wise transformation paradigm to allow for the consistent handling of the data across the various models. This would allow for the smooth integration of the preprocessing steps with the training of the respective models. The preprocessed feature matrix was later used for the training of the ANN-based, CNN-based, and LSTM-based models.

Figure 2 shows the Distribution of Clinical, Demographic, and Diagnostic Variables in the Cleaned Dataset The continuous variables, such as age, systolic and diastolic blood pressure, pulse rate, weight, and SpO_2, follow a near-unimodal distribution with moderate dispersion. The age variable ranges from young adulthood to over 90 years, while blood pressure, pulse rate, and weight are restricted to biologically plausible ranges. The SpO_2 variable has very limited dispersion, with most of the data points concentrated at higher saturation levels. The categorical and binary variables, such as sex,

chest pain, cholesterol level, diabetes, hypertension, habits, ECG, ECHO, TMT, CAG, and the target variable, have discrete distributions with mild class imbalance in some categories.

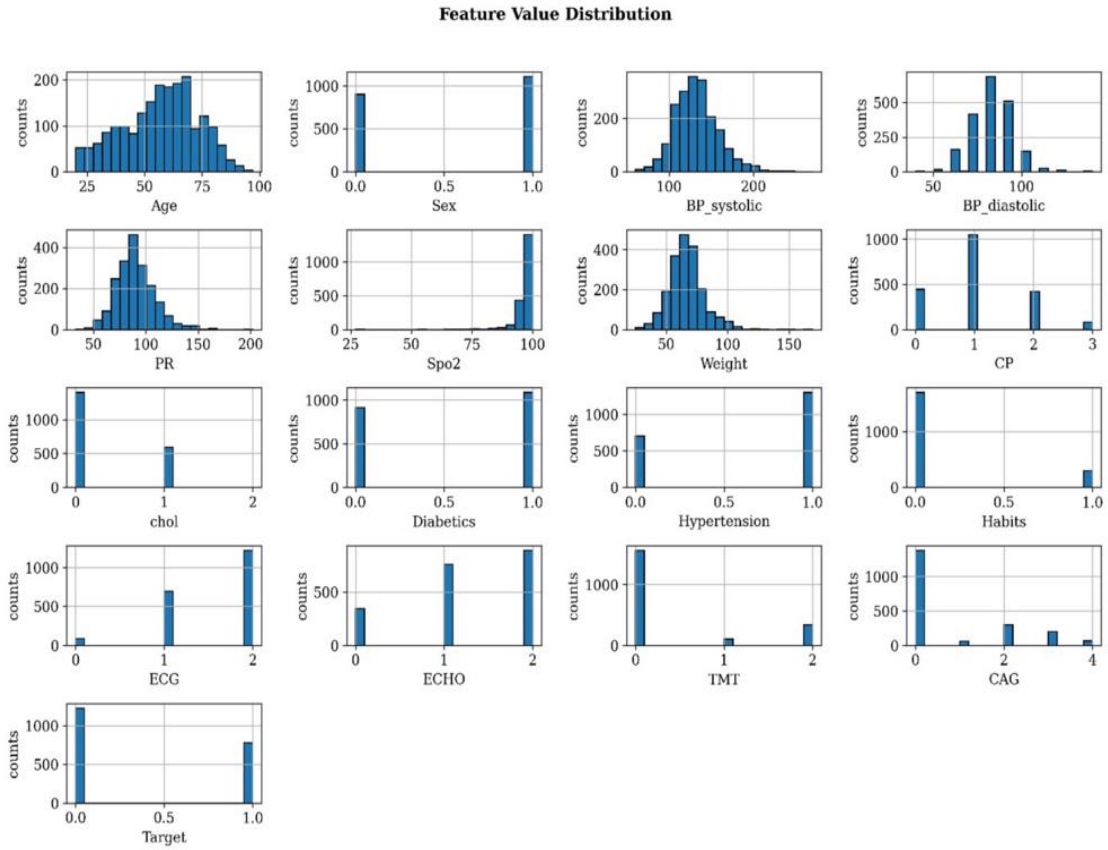

Figure 2: Distribution of Clinical, Demographic, and Diagnostic Feature Values in the Cleaned Dataset

Final Dataset Construction and Splitting. The dataset was split into a training set (80%) and a testing set (20%) using stratified sampling. This is done to maintain class balance. Also, all preprocessing steps were combined into a single pipeline for scikit-learn. In addition, the total number of samples considered, the number of heart disease patients, training samples, and testing samples are provided in Table 1.

Table 1: Total, Training and Testing Samples

Case	Training (80%)	Testing (20%)
Normal	977	244
Heart Disease	623	156
Total	1600	400
Total Samples	2000	

Deep Learning Models

In the proposed research, the deep models of artificial intelligence, including the ANN, CNN, and LSTM models, will be considered to examine their usefulness in the detection of heart disease through clinical results. For the proposed research, the deep models will be considered using the Keras tool.

Artificial Neural Network (ANN)

Generally, in order to predict heart disease, a series of complex, nonlinear relations exist between heterogeneous clinical attributes, such as demographic, physiological, laboratory, and diagnostic test-related variables. Therefore, a fully connected Artificial Neural Network (ANN), as a predictive tool, has been adopted in this study

The proposed ANN accepts a preprocessed clinical feature vector

$\mathbf{x} \in \mathbb{R}^d$ (1)

where d = total number of encoded input attributes obtained from patient data. The various attributes are: systolic and diastolic blood pressures, pulse rate, age, oxygen saturation, weight, sex, chest pain type, cholesterol level, diabetic and hypertensive status, lifestyle habits, and diagnostic outcomes from ECG, ECHO, TMT, and CAG examinations [9]. The proposed ANN jointly learns nonlinear interactions across all clinical attributes, making it well-suited for robust heart disease classification.

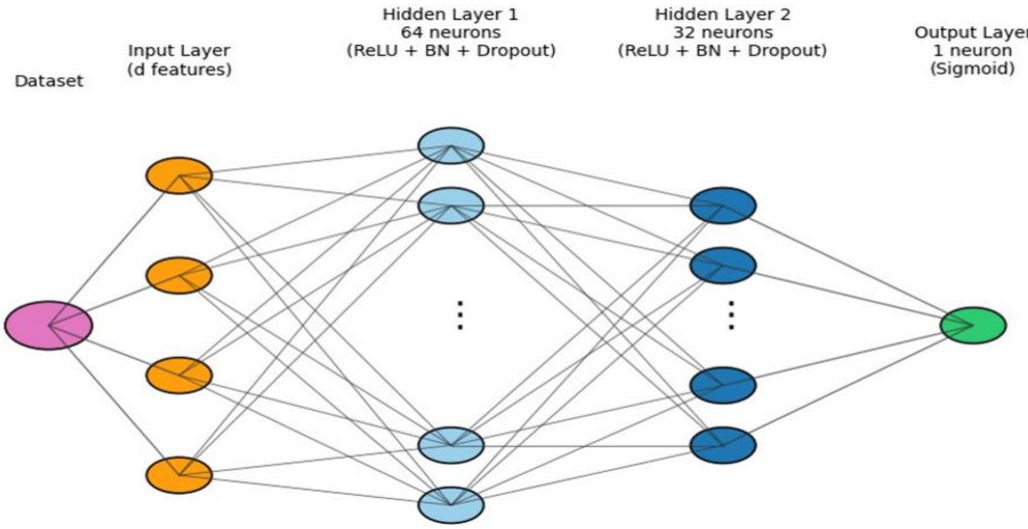

Figure 3: Representative neuron-level architecture of Proposed ANN with two hidden layers, batch normalization, dropout regularization, and sigmoid-based binary output.

The figure 3 shows that the network architecture consists of an architecture of two hidden layers with 64 and 32 neurons, respectively. Each hidden layer applies an affine transformation, followed by batch normalization and ReLU for activation. Batch normalization has been used to reduce problems of internal covariate shift and stabilize training, while dropout regularization with a rate of 0.3 has been used to avoid overfitting by randomly turning off neurons during training.

Mathematically, the transformation occurring at each hidden layer l is given by:

$$\mathbf{a}^{(l)} = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}))(2)$$

where $\mathbf{W}^{(l)}$ and $\mathbf{a}^{(l-1)}$ denote the weight matrix and bias vector of the -th layer, respectively, and $a^{(l-1)}$ represents the activation from the previous layer.

The output layer consists of a single neuron with a sigmoid activation function, which produces a probabilistic estimate of heart disease presence:

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (3)$$

y belongs to 0 and 1 which represents a predictive probability of positive class.

The network is trained using the binary cross-entropy loss function:

$$\mathcal{L}(y, \hat{y}) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \quad (4)$$

where y€ {0,1} denotes the ground-truth class label. and optimized using the Adam optimizer with a learning rate of 10^{-3}, enabling adaptive gradient-based learning and efficient convergence.

Convolutional Neural Network (CNN)

Heart disease continues to be among the primary causes of morbidity and mortality worldwide. It demands effective prediction models to aid in early diagnosis and application during decision-making procedures. Herein, this study utilized a convolutional neural network to identify local interactions among the data features represented as tabular formats. The CNN employs clinical attributes like age, systolic as well as diastolic blood pressures, pulse rates, oxygen saturation, weight, sex, type of chest pains, cholesterol, diabetic, hypertensive, lifestyle, and the results of ECG, ECHO, TMT, and CAG tests. Now, let the pre-processed feature vector be denoted as follows:

$$\mathbf{x} \in \mathbb{R}^d \quad (5)$$

To enable convolutional processing, the input vector is reshaped into a one-dimensional tensor:

$$\mathbf{X} \in \mathbb{R}^{d\times 1} \quad (6)$$

The reshaped input is fed to two one-dimensional convolutional layers with 64 and 128 filters, respectively, using a kernel size of 3 and same padding. Each convolutional layer showed in figure 4 is followed by Batch Normalization (BN), ReLU activation, and max-pooling in order to extract discriminative feature representations while reducing dimensionality. Such feature maps are flattened and fed into a fully connected layer of 64 neurons with ReLU activation, followed by dropout regularization with a rate of 0.3.

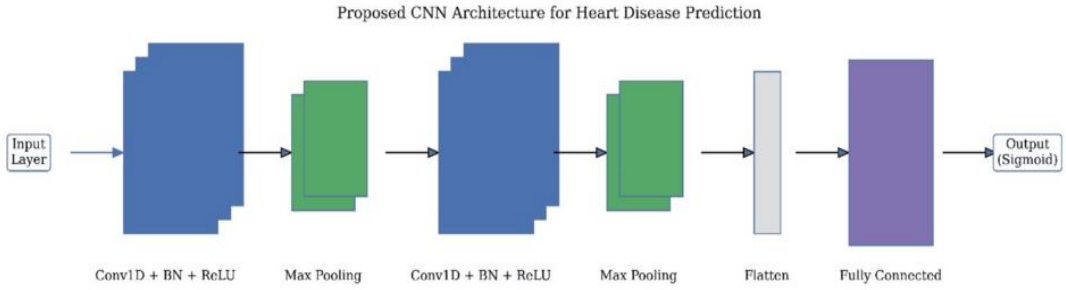

Figure 4: CNN architecture used for heart disease prediction

Finally, the output layer consists of one neuron with a sigmoid activation function to account for the probability of heart disease presence:

$$\hat{y} = \frac{1}{1+e^{-z}} \quad (7)$$

The CNN is trained by minimizing a binary cross-entropy loss function:

$$\mathcal{L}(y, \hat{y}) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \quad (8)$$

where y belongs to 0 and 1 represents the ground-truth label.

Finally, optimization is performed using the Adam optimizer with learning rate:

$$\alpha = 10^{-3} \quad (9)$$

The proposed CNN effectively captures localized feature dependencies and nonlinear relationships by reshaping clinical features and applying convolution operations, thus improving heart disease classification performance.

Long Short-Term Memory (LSTM)

To incorporate the contextual as well as the long-range dependencies among the clinical attributes, a Long Short-Term Memory (LSTM) scheme is adopted, where the entire clinical feature set is taken as a one-dimensional sequence. Here, the model uses the patient attributes like age, systolic blood pressure, diastolic blood pressure, pulse rate, SpO2, weights, sex, type of chest pain (CP), cholesterol level (chol), diabetic status, hypertension status, lifestyle habits, and

output values like ECG, ECHO, TMT, CAG, etc. Though the nature of the input data is tabular, the entire clinical attributes taken as a one-dimensional sequence make the LSTM model efficient in capturing the contextual inter-feature dependencies.

Let the pre-processed clinical feature vector be denoted as:

$\mathbf{x} \in \mathbb{R}^d$ (10)

The feature vector is reshaped into a sequential representation:

$\mathbf{X} = \{x_1, x_2, \ldots, x_d\}$ (11)

where each clinical attribute is treated similarly as a timestep in a sequence. Batch normalization is used before the sequence model for stability.

The figure 5 illustrates that the normalized sequence is fed to a single LSTM layer of size 64 to effectively model long-range feature dependencies. To prevent overfitting, the model makes use of dropout regularization. The last hidden state is then sent to the output layer, which is sigmoid-activated to determine the probability of the presence of heart disease:

$\hat{y} = \frac{1}{1+e^{-z}}$ (12)

The accuracy of proposed LSTM model is enhanced due to use of binary cross entropy loss function. Additionally, the model is tuned using the Adam optimizer with a set learning rate $\alpha=10^{-3}$. This model efficiently handles the contextual association of non-linearly involved heterogeneous clinical attributes.

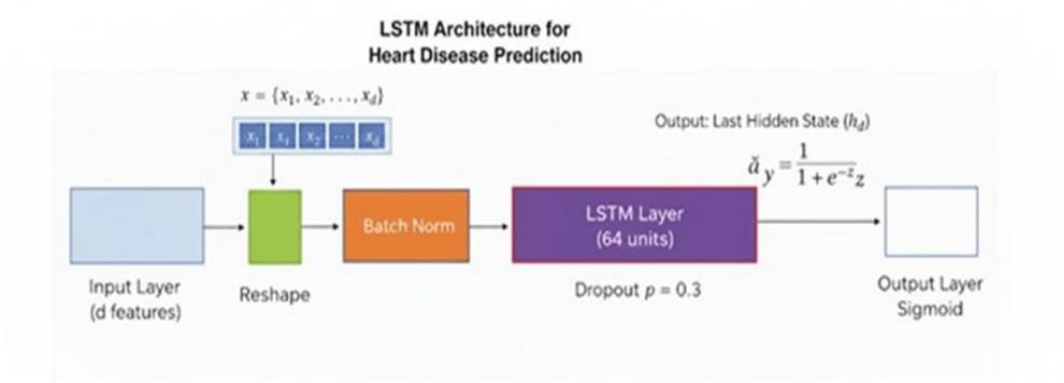

Figure 5: LSTM architecture used for heart disease prediction

Furthermore, all three models were trained with the Adam optimizer and binary cross-entropy loss. Architectural design and training configurations for all the models were consistent in order to perform an unbiased comparative evaluation.

Hyper Tuning

All the three proposed models were developed and implemented with the same set of training strategies to ensure that there was a fair and unbiased comparison. The dataset was then divided into data for training and data for testing using a stratified hold-out method, with 80% for training and the remaining 20% for testing.

***Adam optimization**:* The training of the models was carried out using the Adam optimization technique with a constant rate of 10^{-3}.

***Binary cross-entropy**:* The optimization technique was chosen based on efficiency and robustness in the training of deep neural models. Binary cross-entropy was selected as the evaluation metric since the problem at hand involves a binary classification. The models are trained until a maximum of 100 epochs with a batch size of 16.

***Early Stopping**:* To prevent overfitting and improve generalization capabilities of the model, an early stopping strategy involving validation loss was adopted, with patience equal to 15 epochs and automatic restoring of best model weights.

***Learning rate schedule**:* Moreover, a learning rate adjusting strategy was adopted based on a reduce "on plateau" strategy that

would reduce the learning rate by a factor of 0.5 when validation loss would plateau for five consecutive epochs. In addition to this, 20 percent of the original training set would be held back as validation during training.

To further enhance robustness and evaluate the generalization performance, stratified five-fold cross-validation was further performed. The performance measures were averaged across folds and receiver operating characteristic (ROC) curves were constructed to evaluate the performance of the classifier at different decision levels.

Evaluation Criteria

Several evaluation criteria were taken into consideration for the overall performance evaluation of the proposed models, which not only included the criteria of classification accuracy but also the clinical reliability.

Accuracy was employed to determine the overall number of correctly classified instances. Precision and recall were employed to determine the correctness of positive predictions and the ability to determine true positive instances, respectively. The accuracy is measured by Eq. (13), where TP is true-positive, FP is false-positive, FN is false negative, and TN is true-negative.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

Precision is determined by Eq. (14), which shows the ratio of the number of correct positive predictions to the total number of predictions that the model made as positive, or the level of trustworthiness of the positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

Recall is evaluated using Eq. (15), which refers to the proportion of correct positive responses that the model has made, or what the model can identify.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (15)$$

The F1-score, which is the harmonic mean of precision and recall, was used to achieve a balanced evaluation of classification, particularly in situations where there is class imbalance. The F1-score is calculated using Eq. (16), which is the harmonic mean of two evaluation matrix, precision and recall, which are equally weighted.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \quad (16)$$

These evaluation parameters provide a comprehensive evaluation of classification performance, particularly under the class imbalance.

Specificity is able to correctly classify negative cases that is patients not having the disease. The significance of this metric in the medical application is very important since it is able to determine the efficiency of the model in preventing false, since healthy patients would not be labeled as diseased. Specificity is defined as TNs/TNs+ FPs as in the equation below (17).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (17)$$

Each of the evaluation criteria was performed using standardized Scikit-learn functions to ensure a high degree of reproducibility and consistency. MCC was included as a robust metric since it takes into consideration all the factors in the confusion matrix and provides a balanced evaluation for situations with unbalanced class distributions. Another metric applied was the receiver operating characteristic curve, which allowed an analysis of model performance across various thresholds for decisions; accordingly, the area under the ROC curve was calculated, which is a measure of the discriminative power of the model irrespective of the decision threshold.

For hold-out evaluation, metrics were computed on the test set, while stratified 5-fold cross validation was used to assess robustness and generalization. This combination of metrics and evaluation strategies ensures a reliable and clinically meaningful assessment of model performance.

RESULT

Feature Importance

In this proposed model, in order to depict the structural nature of the dataset as well as the predictive nature of the models, following bar chart is plotted, as shown in Figure 6, The bar chart displays feature

importance scores from a Random Forest model - higher values indicated greater influence on predictions. "ECHO" has the highest importance, it means it is contributed most to the model's decisions, followed by "CAG" and "ECG". Lower bars like "Sex," "Habits," and "Diabetics" show these features had minimal impact on the model's performance.

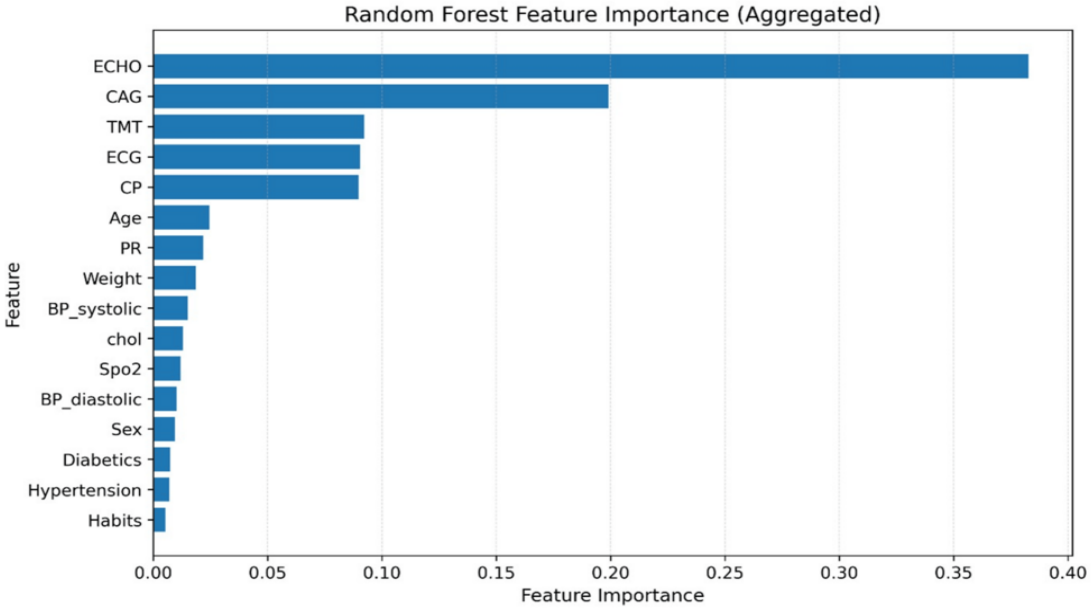

Figure 6: Important Features identified by Random Forest

Random Forest computes importance based on how much each feature reduces impurity across all trees in the forest, it has been observed that there exist significant correlations between multiple feature variables, thus creating a scenario of multicollinearity associated with clinical as well as diagnostic variables.

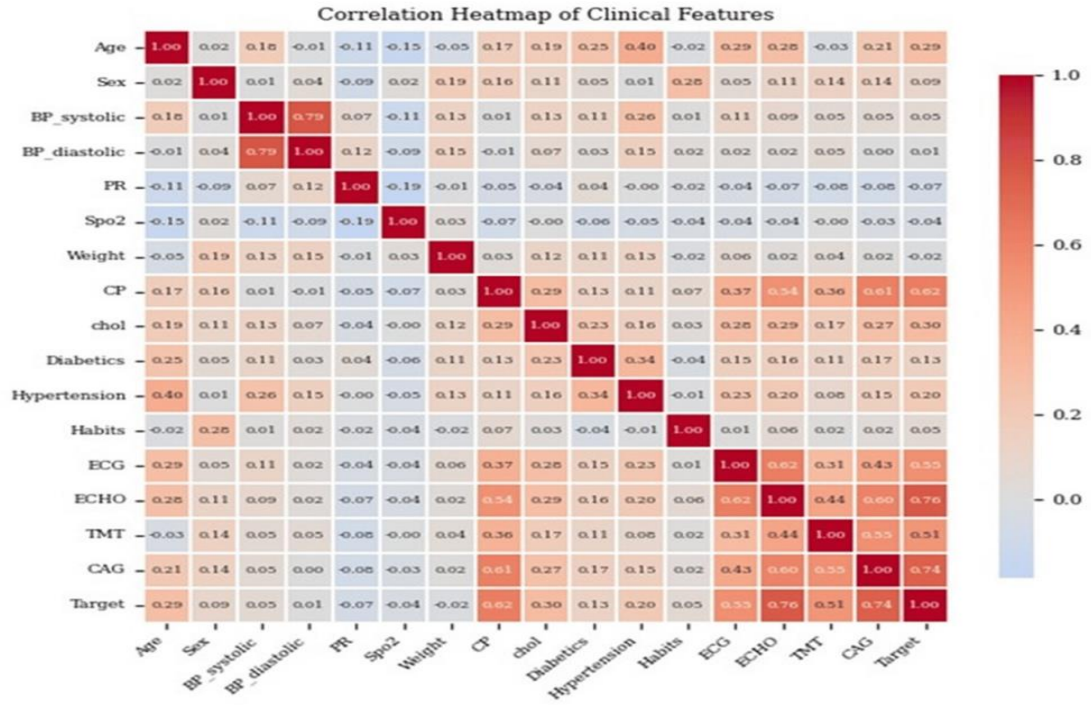

Figure 7: Correlation of different features through Heatmap

The correlation heatmap of the clean clinical data set, as shown in Figure 7, reveals that the diagnostic variables ECG, ECHO, TMT, and CAG have a high positive correlation rate with the heart disease target, as shown by the correlation coefficients $|r| \approx 0.55\text{-}0.75$. Likewise, the clinical risk-related variables chest pain, cholesterol level, hypertension, diabetes, and age have moderate positive correlation rates with the heart disease target, as shown by the correlation coefficients $|r| \approx 0.20\text{-}0.40$. Conversely, demographic and physiologic variables such as sex, weight, pulse rate, and SpO2 have low and near-zero correlation rates with the heart disease target, as shown by the correlation coefficients $|r| < 0.10$. In this data set, there is multicollinearity among the blood pressure variables, namely systolic and diastolic blood pressure. Nevertheless, as mentioned in the previous section, the proposed algorithm is capable of handling multicollinearity and is therefore suitable.

Figure 8 illustrates the detailed performance capability of the ANN, CNN, and LSTM models, along with the training and validation loss, as well as the test set confusion matrices. The converting model has a stable convergence process and a stable reduction in the loss that can be noticed through the epochs, indicating successful learning in the ANN, CNN, and LSTM models. The ANN model has a fast convergence process and reaches a low value of loss, but there is a slight variation between the training and validation loss in the last epochs, indicating a mild overfitting process. The CNN model indicates a little higher variation between the training and validation loss, which indicates a relatively poor generalization capability. On the other hand, the LSTM model has a parallel training and validation loss process with a smooth convergence, indicating a stable learning process with reduced overall predictive accuracy.

These are further confirmed by the confusion matrices. The ANN model has the most balanced performance, and it has correctly identified 242 instances of normal and 147 instances of disease with only 2 false positives and 9 false negatives. The CNN model is also performing well with 240 normal and 151 disease instances correctly identified with a slightly higher level of misclassification. The LSTM model has a higher false positive rate (13 normal instances missed) and lower accuracy. Taken together, these results indicate that the ANN model has the highest sensitivity and specificity followed by the CNN model and that the LSTM model cannot be used with structured tabular clinical data.

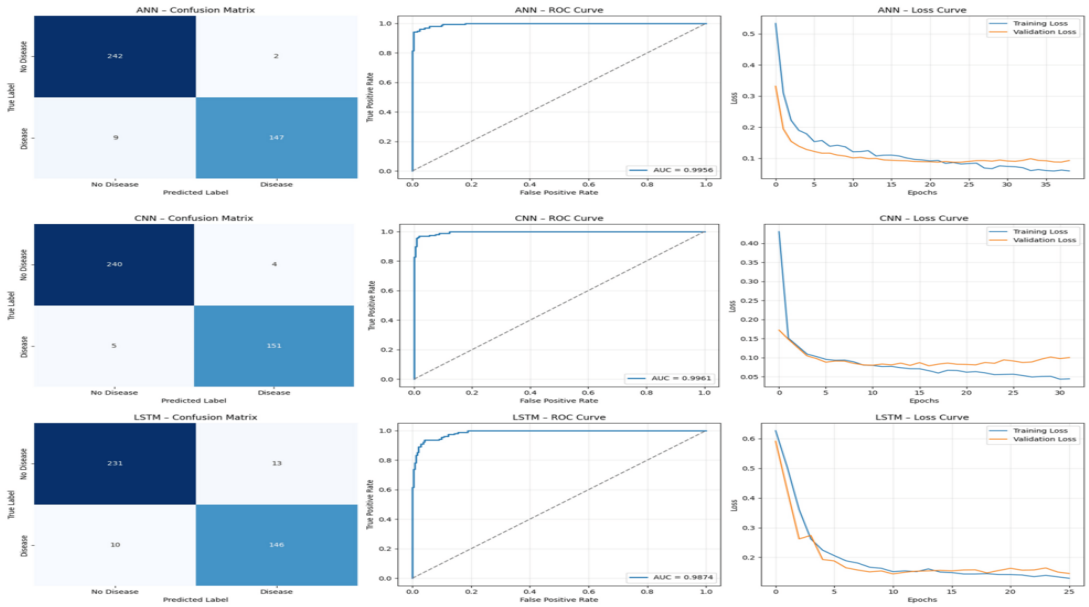

Figure 8: Training- validation loss and, confusion matrices of ANN, CNN and LSTM Models.

Cross-Validation Analysis

To measure model robustness, stratified five-fold cross-validation has been performed. The ROC curves in Table 2, which measure cross-validated discriminative performance, show that the selected models have performed consistently. Both ANN and CNN models have retained ROC - AUC values corresponding to a high discriminative capacity, greater than 0.99, while in the case of the LSTM model, a lower, yet high, ROC - AUC has been observed. The fact that performance trends are being observed here implies that the performance does not come from a training-test split, hence increasing its reliability.

Table 2: Cross-validated performance for ANN, CNN, and LSTM models used for heart disease prediction

Model	Validation Accuracy	Validation AUC	Test Accuracy	Test AUC	Precision	Recall	F1-score	MCC
ANN	0.963	0.993	0.962	0.992	0.950	0.951	0.950	0.919
CNN	0.958	0.991	0.951	0.990	0.939	0.935	0.936	0.896
LSTM	0.938	0.980	0.932	0.978	0.901	0.928	0.914	0.858

Performance analysis of proposed work

This section focuses on the presentation of the evaluation of the proposed deep learning models, with a comparative analysis of the effectiveness of the proposed approach in the detection of cardiovascular diseases like early heart disease. ANN, CNN, and LSTM deep learning models are used in the evaluation process, with the help of a stratified hold-out test set followed by the cross-validation technique.

Table 3: Performance analysis of ANN, CNN, and LSTM models on the hold-out test set for heart disease prediction

Model	Accuracy	Precision	Recall	F1-score	Specificity	MCC	AUC
ANN	0.9620	0.946179	0.957659	0.951539	0.964788	0.920788	0.993566
CNN	0.9565	0.956367	0.931993	0.943326	0.972175	0.909097	0.992981
LSTM	0.9330	0.930876	0.896005	0.911335	0.956587	0.860021	0.984466

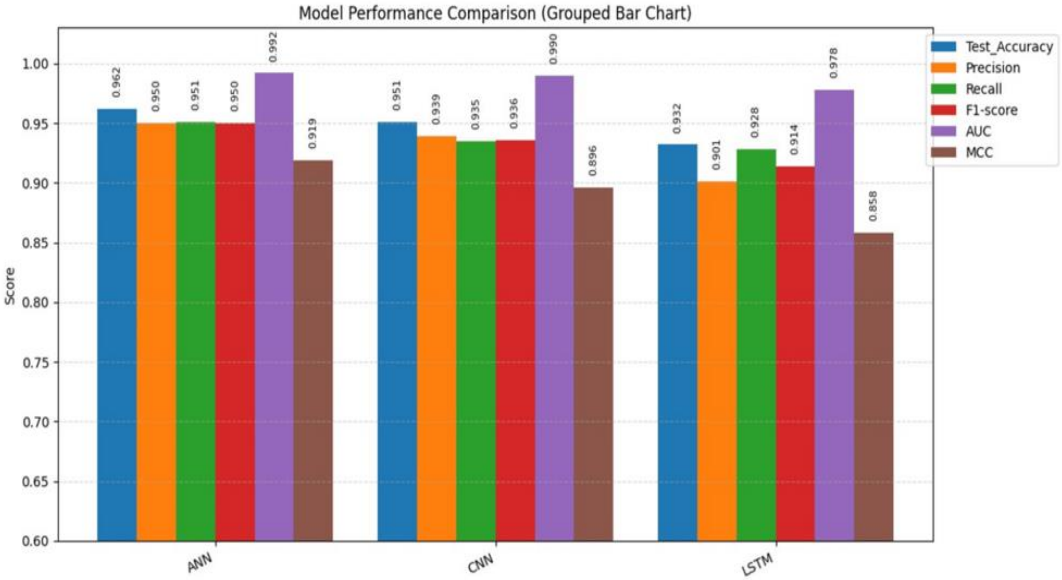

Figure 9: Performance analysis of ANN, CNN, and LSTM

The stratified hold-out testing and stratified five-fold cross-validation are employed to evaluate the testing performance of the proposed ANN, CNN, and LSTM models in terms of the clinically relevant metrics. Table 3 and figure 9 illustrate that the ANN model

performs the best in terms of overall test performance, accuracy (96.20%), recall (0.9577), MCC (0.9208), and AUC (0.9936), indicating that the ANN model has high discriminative power and balance.

The CNN model has similar test accuracy of 95.65% and the highest specificity of 97.22% and can better distinguish the non-heart disease samples with less false positives, but with higher computational complexity. Compared with the other models, the LSTM model has lower accuracy of 93.30%, AUC of 0.9845, which could be explained by the fact that the tabular clinical information is static, but the recall value may indicate its potential value in a longitudinal study.

Generally, the results confirm that ANN is the best model to be applied in practice for predicting heart diseases with the optimal accuracy, robustness, and efficiency ratio, while CNN and LSTM can also be considered as good alternatives depending on the requirements of the task.

DISCUSSION

The current paper has conducted a systematic comparison of Artificial Neural Network (ANN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) networks to predict heart disease based on a multimodal clinical dataset of Ayush Multi Speciality Hospital and Research Centre. The findings indicate that the three architectures had high rates of predictive performance as well, which substantiates the idea that deep learning models will be capable of assisting in early cardiovascular risk prediction when trained on clinically enriched structured data [10,12].

The key advantage of the current study is the inclusion of sophisticated diagnostic variables like ECG, ECHO, TMT, and CAG along with demographic and physiological characteristics. The analysis of feature importance revealed ECHO and CAG to be the most effective predictors, as opposed to ECG, which suggests that structural and functional cardiac tests can be a good source of information to use in the predetermination of the disease. The same has been reported in previous cardiovascular prediction studies whereby diagnostic imaging, as well as stress-related cardiac parameters, has significantly enhanced predictive discrimination [13,18]. Correlation study also revealed that there were strong and positive correlations between these diagnostic variables and the heart disease target which supports the usefulness of multimodal diagnostic integration.

Despite the fact that CNN slightly performed better at validation accuracy during stratified five-fold cross-validation, ANN showed the largest balanced overall performance at validation and independent testing. Specifically, ANN performed better in terms of recall, MCC, and ROC-AUC, which implies that it is more sensitive and more dependable in terms of class-balanced discrimination. In clinical decision-support systems, recall is of especial concern because false negative outcomes may cause treatment and follow-up patient risk.. The elevated MCC also attests to the fact that ANN was not subject to severe imbalance in regards to classes. The higher ANN performance may be attributed to the data organization of the input data. The current data is composed of heterogeneous clinical variables in tabular format as opposed to spatially ordered signals. ANN architectures are full-connected and are suitable in learning nonlinear interactions of such variables without any local spatial assumptions [14,16].

Compared to this, CNN requires converting tabular data into one-dimensional convolutional features, which can synthetically form neighbourhood relationships that do not always represent clinical relationships in full. The CNN was, however, very competitive, which implies that some localized feature interactions are still involved in prediction. The predictive accuracy of LSTM was relatively lower than ANN and CNN, but its ROC-AUC was also high. This is not surprising given that LSTM models are mostly intended to learn temporal relations among consecutive observations

[15,20]. In the current project, the ordered sequences of static tabular features were used, which restricted the usefulness of the recurrent memory mechanisms. Nevertheless, the consistent convergence at the time of training indicates that LSTM can be more beneficial in the case of longitudinal patient records, follow-up measurements repeated over a time period, or wearable cardiac monitoring data. These findings are further supported by analysis of confusion matrix. ANN was the lowest in the combination of false positive and false negative rates, which meant that it correctly identified both diseased and non-diseased patients. CNN had a slightly better specificity, implying better negative-case recognition whereas LSTM had a relatively better misclassification. These findings reveal ANN to provide the best balance to structured clinical classification.

The current work is superior to previous investigations, which relied primarily on the benchmark repositories like the UCI and Kaggle datasets [16,17], since the current work is characterized by real-time hospital-generated data that include clinically richer diagnostic characteristics. Moreover, MCC, specificity, recall and ROC-AUC utilization offer a more clinically significant assessment in comparison to accuracy [19]. Stratified hold-out testing coupled with stratified five-fold cross-validation thus enhances faith in robustness of the models and less reliance on one split of the data. Regarding the practicality of ANN, it has a significant deployment benefit due to its ability to achieve a high predictive accuracy with a lower architectural complexity and lower computation cost than CNN and LSTM. This is why ANN would be especially applicable to hospital-based decision-support systems where quick inferences and reproducibility are important.

In spite of these strengths, there are a number of limitations. The data set was collected in one institution and therefore this could be limiting the applicability to wider population. Class imbalance and moderate gender imbalance can also be a factor of model learning. Moreover, despite the fact that the importance of features of the Rand Forest increased the interpretation, the deep learning decision paths are not transparent enough to be directly explicable by the clinician.

Further studies ought to be directed at multicentre validation, increase in the number of patients to study and incorporation of Explainable Artificial Intelligence techniques to enhance transparency and confidence by clinicians. The ANN architectures feature-fusion, multimodal attention, and hybrid deep learning can be further predictive-robust, particularly when they are used together with longitudinal cardiovascular monitoring data.

All in all, the results suggest that ANN is the most suitable strike between predictive accuracy, robustness, and computational efficiency to predict structured multimodal heart diseases at the moment, whereas CNN and LSTM can be considered an option in the future, depending on the characteristics of the new data and the needs of clinical applications.

CONCLUSION

This paper is able to prove that deep learning-based models can be highly effective in predicting heart disease using a rich real-life clinical dataset over AMSHRC Pvt. Ltd., Vijayapura, Karnataka, India. Cardiovascular disease is an important condition that requires timely clinical intervention and proper management of risks and this can only be achieved through early and accurate diagnosis of the condition as the conventional methods of diagnosis can be difficult to capture nonlinear trends in heterogeneous patient records. The proposed framework is a unified system of preprocessing, feature extraction, and training of three types of neural networks ANN, CNN, and LSTM, with the use of structured clinical data, and allows a fair and complete comparison of their predictive quality.

Additional advanced diagnostic characteristics including ECG, ECHO, TMT

and CAG, along with the traditional demographic and physiological variables, enabled the models to be trained on a more comprehensive and more clinically meaningful feature space than most publicly available datasets. The robustness was ensured with stratified five-fold cross-validation and independent hold-out test evaluations which were implemented to control sampling bias and validate their generalizations performance. Although CNN had the best validation accuracy and MCC, proving to be effective at extracting a hierarchical feature representation, the ANN model was the most well rounded and effective on the whole. ANN had good accuracy of 96.20 and ROC -AUC of 0.9936 with high recall and MCC scores on the test set, which makes it useful in clinical where binary classification is required and it is very important to both differentiate between diseased and non-diseased cases. In future this work will explore the Feature-Fusion Artificial Neural Network (FF-ANN) and the integration of explainable AI approaches to it.

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. World Health Organization, "Cardiovascular Diseases (CVDs)" WHO Fact Sheets, Jul. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Zahro R, Arzikulov F. Early Detection And Prognosis Of Chronic Heart Diseases Using Artificial Intelligence. EJHSMI [Internet]. 2026 Jan. 9 [cited 2026 May 19];2(1):53-64. Available from: https://eurekaoa.com/index.php/5/article/view/140
3. Michalowski M, Sun-Mitchell S, Delaney CW. Principles of Artificial Intelligence and Big Data in Healthcare. InBridging Artificial and Human Intelligence: Implementation Strategies and Case Studies in Healthcare 2026 Jan 13 (pp. 25-41). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-11938-4_2
4. Gul G, Korejo IA, Hakro DN, Alqahtani H, Abbasi A, Babar M, Al Rahbi O, Ali NI. Machine Learning and Ensemble Methods for Cardiovascular Disease Prediction: A Systematic Review of Approaches, Performance Trends, and Research Challenges. *Computers*. 2026; 15(1):25. https://doi.org/10.3390/computers15010025
5. Chibueze KI, Didiugwu AF, Ezeji NG, Ugwu NV. A CNN based model for heart disease detection. Scientia Africana. 2024 Aug;23(3):429-42. DOI:10.4314/sa.v23i3.38
6. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677. 2015 Nov 11.
7. Diallo, R., Edalo, C., Awe, O.O. (2025). Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In: Awe, O.O., A. Vance, E. (eds) Practical Statistical Learning and Data Science Methods. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health. Springer, Cham. https://doi.org/10.1007/978-3-031-72215-8_12
8. Hughes A, Shandhi MM, Master H, Dunn J, Brittain E. Wearable devices in cardiovascular medicine. Circulation research. 2023 Mar 3;132(5):652-70. doi: 10.1161/CIRCRESAHA.122.322389.
9. Budhal, Sunanda. (2025). Heart Disease Prediction Using Machine Learning On A Real-Time Clinical Dataset With Multi-Diagnostic Features. International Journal of Applied Mathematics. 38. 2719-2742. 10.12732/ijam.v38i12s.1584.
10. Kumar R, Garg S, Kaur R, Johar MGM, Singh S, Menon SV, Kumar P, Hadi AM, Hasson SA, Lozanović J. A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. Front Artif Intell. 2025 May 13; 8:1583459. doi: 10.3389/frai.2025.1583459.
11. Acharya, U.R., Fujita, H., Oh, S.L. *et al.* Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Appl Intell* 49, 16–27 (2019). https://doi.org/10.1007/s10489-018-1179-1

12. Elyamani, H.A., Salem, M.A., Melgani, F. *et al.* Deep residual 2D convolutional neural network for cardiovascular disease classification. *Sci Rep* 14, 22040 (2024). https://doi.org/10.1038/s41598-024-72382-3
13. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018 Oct 1;25(10):1419-1428. doi: 10.1093/jamia/ocy068.
14. Manimaran G, Peimankar A, Puthusserypady S, Momeni M, Asyari RAI, Jahan MS, Moll J, Wiil UK, Ebrahimi A. Explainable deep learning based techniques for ECG-Based heart disease classification: A systematic literature review and future direction. Comput Biol Med. 2025 Dec; 199:111324. doi: 10.1016/j.compbiomed.2025.111324.
15. Banerjee T, Paçal İ. A systematic review of machine learning in heart disease prediction. Turk J Biol. 2025 Sep 11;49(5):600-634. doi: 10.55730/1300-0152.2766.
16. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [Dataset]. UCI Machine Learning Repository. 1989. URL https://archive. ics. uci. edu/ml/datasets/heart+ Disease.
17. The Devastator: "Predicting heart disease risk using clinical variables,". Kaggle Dataset. 2025, https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var
18. Shinkins B, Yang Y, Abel L, Fanshawe TR. Evidence synthesis to inform model-based cost-effectiveness evaluations of diagnostic tests: a methodological review of health technology assessments. BMC Med Res Methodol. 2017 Apr 14;17(1):56. doi: 10.1186/s12874-017-0331-7.
19. Di Cesare M, Perel P, Taylor S, Kabudula C, Bixby H, Gaziano TA, McGhie DV, Mwangi J, Pervan B, Narula J, Pineiro D, Pinto FJ. The Heart of the World. Glob Heart. 2024 Jan 25;19(1):11. doi: 10.5334/gh.1288.
20. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications. 2018 May;29(10):685-93. DOI:10.1007/s00521-016-2604-1

How to cite this article: Sunanda Budhal, Sheetalrani Kawale, Bhagirathi Hallalli, Nitin Agarwal. Systematic benchmarking of ANN, CNN, and LSTM models for heart disease prediction using multimodal clinical features. *Int J Health Sci Res.* 2026; 16(5):176-190. DOI: *https://doi.org/10.52403/ijhsr.20260521*

******