Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Evaluation of BiLSTM-CNN, XGBoost, and Ridge Regression for Heart Disease Classification on the Cleveland Dataset

Ajimah Nnabueze Edmund[1], Ikiomoye Douglas Emmanuel[2], Esenogho Ebenezer[3]

[123]Centre for Artificial Intelligence and Multidisciplinary Innovations, Department of Auditing, College of Accounting Sciences, University of South Africa, Pretoria 0002, South Africa.

Received:
October 15, 2025
Revised:
May 17, 2026
Accepted:
June 25, 2026
Published:
June 27, 2026

Corresponding Author:
Author Name*:
Ikiomoye D. Emmanuel
Email*:
emmanid@unisa.ac.za

DOI:
10.63158/journalisi.v8i3.1668

Abstract. Transformers have become the dominant architecture for tabular data modelling in natural language processing; however, their effectiveness for numerical tabular classification on modest sized and moderately imbalanced datasets remains unclear. This study evaluates the performance of hybrid deep learning and classical machine learning models which use the Cleveland Heart Disease dataset with 297 complete observations and was artificially constructed from 13 clinical features. The models examined include BiLSTM-CNN, Random Forest, XGBoost, Logistic Regression, and Ridge Regression. An experimental comparative approach was adopted under identical preprocessing, training conditions, and evaluation metrics, including accuracy, recall, F1-score, and Area Under the Curve (AUC). Results show that BiLSTM-CNN achieved the highest recall (0.8478), demonstrating strong minority class detection capability. Random Forest and XGBoost produced the best-balanced performance with 81.67% accuracy and the BiLSTM-CNN has the best F1-score of 0.8364, while Ridge Regression achieved the highest AUC (0.8945). This study provides empirical evidence that hybrid recurrent and ensemble models perform optimally on a small to medium sized Cleveland Heart Desease numerical tabular datasets without pre-training, offering practical guidance for Cleveland Heart Disease tabular clinical classification tasks, and no external validation was performed.

Keywords: BiLSTM-CNN, Random Forest, heart disease classification, imbalanced learning, AUC, Small Medical

1. INTRODUCTION

Since heart disease is still one of the world's top causes of death, early and precise diagnosis is essential for clinical intervention. Longitudinal vital signs and ECG recordings are examples of sequential patient data that reflect temporal patterns suggestive of diseased conditions. It is uncertain, therefore, whether model architecture is best for categorizing such numerical sequences, especially on small medical datasets [1]. This work uses the Cleveland Heart Disease dataset (303 samples, 13 clinical characteristics) to compare hybrid BiLSTM-CNN models, and some other conventional classifiers. We assess whether deep sequential models perform better than conventional techniques in this small-data regime, offering useful recommendations for cardiovascular risk assessment. The Cleveland Heart Disease dataset's small size (303 samples), mixed numerical/categorical features, and a moderate class imbalance make it perfect for testing BiLSTM-CNN, XGBoost, and Ridge Regression. Within a clinically relevant binary classification challenge, BiLSTM-CNN investigates sequential feature ordering, Ridge Regression offers a regularized linear baseline, and XGBoost captures non-linear interactions. CNNs can capture local interactions and BiLSTM can describe bidirectional relationships by reshaping tabular features as a sequence, which reveals non-linear patterns. In many real-world applications, the available data are highly imbalanced, where the minority class contains significantly fewer samples than the majority class [2], [3]. No work has rigorously tested BiLSTM-CNN (considering features as sequences) against XGBoost and Ridge Regression under identical pre-processing, evaluation, and class-handling techniques, despite the predominance of small clinical datasets with significant imbalance [1], [4], [5], [6], [7].

Despite the success of deep learning methods, some classical machine learning algorithms remain highly relevant because of their interpretability, computational efficiency, and strong predictive performance on structured datasets. Ensemble tree-based methods such as XGBoost have consistently achieved state of the art performance in tabular data classification tasks due to their robustness, regularization capability, and ability to model nonlinear relationships effectively. Similarly, Ridge Regression, which applies L2 regularization to linear classification problems, provides stable probabilistic predictions, reduced overfitting, and resilience against multi-collinearity in high dimensional data Previous studies have shown that these conventional approaches can

remain competitive even when compared with more computationally expensive deep learning models [3], [8].

The contribution is empirical benchmarking and method comparison, not clinical decision-support exploration, due to artificial feature ordering and small sample size. Several studies have investigated imbalance handling techniques such as Synthetic Minority Oversampling Technique (SMOTE), class weighting, cost sensitive learning, and ensemble balancing methods to improve minority class prediction [9], [10].

Although previous research has explored either deep learning architectures or classical machine learning methods independently, there remains limited work providing a direct and fair comparison between hybrid deep learning models and classical algorithms on imbalanced sequential data. Most comparative studies employ different datasets, preprocessing strategies, or imbalance handling techniques, making performance comparisons inconsistent and difficult to generalize. In addition, many studies prioritize accuracy as the primary evaluation metric even though accuracy can be misleading under imbalanced conditions. Metrics such as recall, F1-score, and Area Under the Curve (AUC) provide more meaningful evaluation because they better reflect minority class detection capability and ranking performance [11], [12]. Furthermore, limited studies have specifically examined the comparative performance of BiLSTM-CNN, XGBoost, and Ridge Regression under identical imbalance handling conditions using SMOTE and consistent evaluation protocols. The capability of Ridge Regression to achieve competitive AUC performance on high dimensional sequence data has also received limited attention in the literature [13], [14].

Therefore, this research intends to provide a rigorous empirical comparison of hybrid deep learning and classical machine learning models for imbalanced sequential classification. The study focuses on BiLSTM-CNN, XGBoost, and Ridge Regression as primary models, while additional baseline models including Random Forest and Logistic Regression are also evaluated. The research aims to implement and optimize these models using a common sequential dataset with a moderate class imbalance while applying consistent imbalance handling techniques such as SMOTE and class weighting to ensure fairness in evaluation. The study further seeks to analyze performance using metrics including accuracy, precision, recall, F1-score, and AUC, with particular emphasis

on recall and AUC because of their importance in minority class prediction. In addition, the research evaluates the strengths and weaknesses of each model in terms of classification capability, robustness, ranking performance, and computational efficiency to provide practical recommendations for model selection in cost sensitive sequential classification applications [15], [16], [17]. The objective of this work is to comprehensively benchmark Ridge Regression, XGBoost, and BiLSTM-CNN on the Cleveland Heart Disease dataset. Research questions: (1) Which model for the classification of heart disease maximizes recall and AUC? (2) On small, unbalanced clinical data, do sequential deep learning architectures perform better than classical approaches?.

The contributions of this research are threefold. First, the study provides a direct and fair experimental comparison between hybrid deep learning, ensemble learning, and regularized linear models, specifically BiLSTM-CNN, XGBoost, and Ridge Regression, under identical experimental conditions on imbalanced sequential data. Second, the research offers new insights into the comparative strengths of the evaluated models by demonstrating the recall superiority of BiLSTM-CNN, the balanced classification performance of XGBoost, and the strong ranking capability of Ridge Regression through AUC analysis [11], [12], [18]. This study is basically a comparative analysis of already established models [6], [15]. Tabular features are actually reshaped into a sequence to treat each of the features as a token, which enables attention mechanisms to model cross-feature interactions very powerfully. Below are focused reviews on some related studies on Cleveland Heart Disease classification.

In [19] Kadhim & Radhi in 2023 worked on 'Heart disease classification using ML algorithms' Proposes a heart disease prediction model using Random Forest, SVM, KNN, and Decision Tree experimented on a Cleveland Heart Disease dataset. Random Forest achieved 94.9% accuracy. Hyper parameter tuning via random search further improved accuracy to 95.4%. The study emphasizes data pre-processing, outlier removal, and standard scaling to enhance model performance.

Hutagalung & Andrianingsih [20] in 2026 worked on 'Heart Disease Classification Using Optimized XGBoost and Random Forest with SHAP Explanations' where they compares Random Forest and XGBoost with random search, Bayesian, and PSO optimisation on the Cleveland dataset using nested cross-validation. RF with PSO

achieved the highest ROC-AUC (0.9089) in a Cleveland Heart Disease dataset. SHAP analysis identified *oldpeak*, *ca*, *thal*, and *cp* as key features, improving model interpretability. Differences among top models were not statistically significant.

2. METHODS

The experiments were performed in the MATLAB (2025b) environment and employed the deep learning toolbox. The study investigates imbalanced binary classification using sequential numerical data with the objective of evaluating the comparative performance of hybrid deep learning and classical machine learning models. The primary models examined in this research include Bidirectional Long Short-Term Memory with Convolutional Neural Network (BiLSTM-CNN), Extreme Gradient Boosting (XGBoost), and Ridge Regression. Additional baseline models including Random Forest and Logistic Regression based classifiers were also implemented to provide broader comparative insights [21], [22].

The experimental framework was designed to ensure fairness and reproducibility across all evaluated models. Identical preprocessing procedures, class imbalance handling techniques, and evaluation protocols were applied during training and testing. Model performance was assessed using metrics suitable for classification under moderate class imbalance problems, namely precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics were prioritized because they provide more reliable evaluation of minority class prediction capability than overall accuracy alone, particularly in healthcare related diagnostic systems where false negatives may have serious consequences. Figure 1 shows the Flowchart of the XGBoost, Ridge Regression and BiLSTM Models.

2.1. Dataset Source

The dataset used in this study is the Cleveland Heart Disease dataset obtained from the University of California, Irvine Machine Learning Repository, see Figure 2 the screenshot [23]. The dataset is one of the most widely used benchmark datasets for cardiovascular disease prediction and machine learning based clinical diagnosis. It contains clinical and demographic information collected from patients undergoing coronary angiography at

the Cleveland Clinic Foundation and has been extensively applied in healthcare analytics and intelligent diagnostic system research.

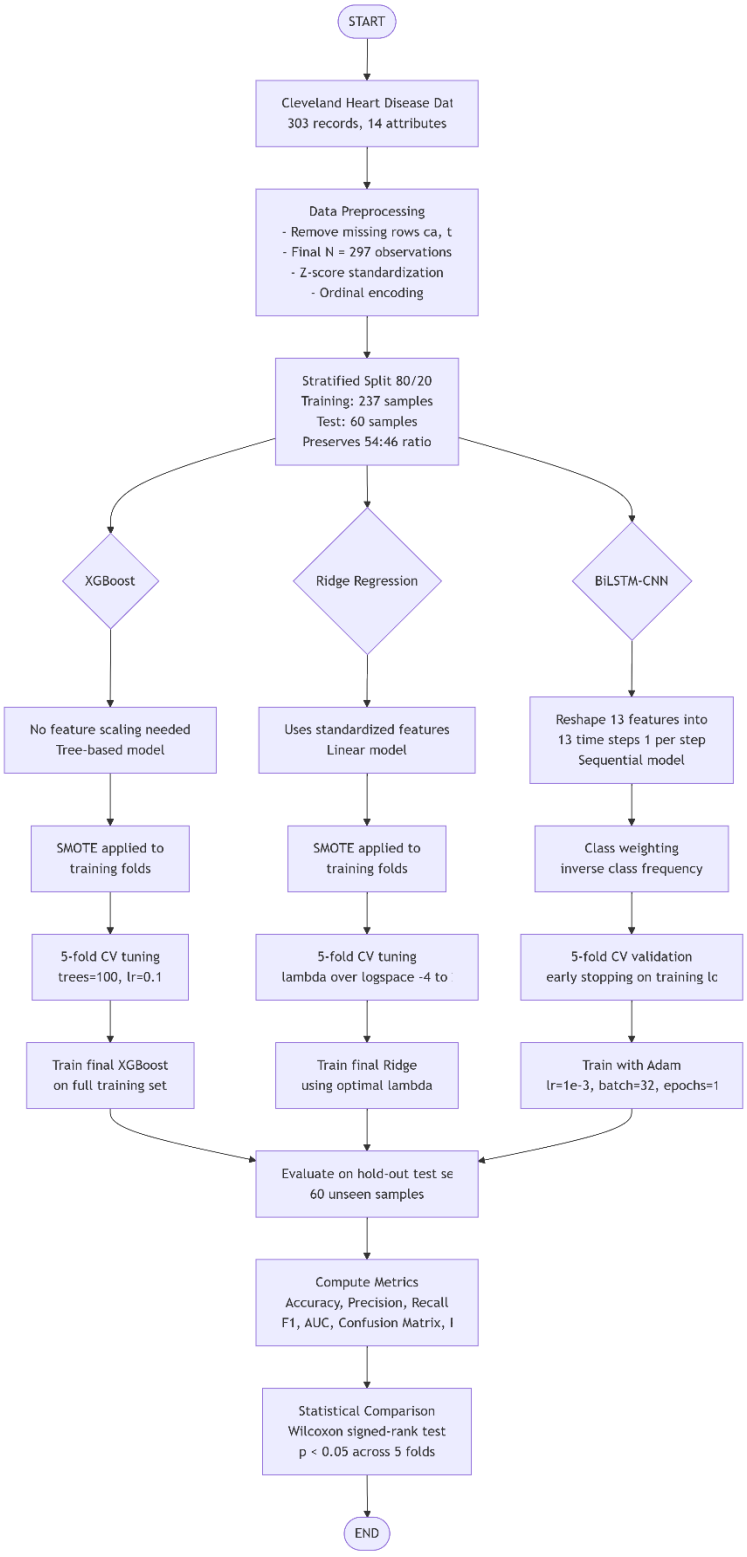

Figure 1. The Flowchart of the XGBoost, Ridge Regression and BiLSTM Models

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	0	145	233	1	2	150	0	2.3	2	0	2	0
3	67	1	3	160	286	0	2	108	1	1.5	1	3	1	1
4	67	1	3	120	229	0	2	129	1	2.6	1	2	3	1
5	37	1	2	130	250	0	0	187	0	3.5	2	0	1	0
6	41	0	1	130	204	0	2	172	0	1.4	0	0	1	0
7	56	1	1	120	236	0	0	178	0	0.8	0	0	1	0
8	62	0	3	140	268	0	2	160	0	3.6	2	2	1	1
9	57	0	3	120	354	0	0	163	1	0.6	0	0	1	0
10	63	1	3	130	254	0	2	147	0	1.4	1	1	3	1
11	53	1	3	140	203	1	2	155	1	3.1	2	0	3	1
12	57	1	3	140	192	0	0	148	0	0.4	1	0	2	0
13	56	0	1	140	294	0	2	153	0	1.3	1	0	1	0
14	56	1	2	130	256	1	2	142	1	0.6	1	1	2	1
15	44	1	1	120	263	0	0	173	0	0	0	0	3	0
16	52	1	2	172	199	1	0	162	0	0.5	0	0	3	0
17	57	1	2	150	168	0	0	174	0	1.6	0	0	1	0
18	48	1	1	110	229	0	0	168	0	1	2	0	3	1
19	54	1	3	140	239	0	0	160	0	1.2	0	0	1	0
20	48	0	2	130	275	0	0	139	0	0.2	0	0	1	0
21	49	1	1	130	266	0	0	171	0	0.6	0	0	1	0
22	64	1	0	110	211	0	2	144	1	1.8	1	0	1	0
23	58	0	0	150	283	1	2	162	0	1	0	0	1	0

Figure 2. A screenshot of the Cleveland Heart Disease dataset obtained from the University of California

The original dataset contains 303 patient records and 14 clinical attributes. Records containing missing values in the *ca* (number of major vessels) and *thal* (thalassemia status) variables were removed during preprocessing, leaving 297 complete instances for analysis. The dataset attributes are grouped as follows:

1) Demographic Features

   Sex: gender of the patient (0 = female, 1 = male), *Age*: age of the patient in years.

2) Clinical and Laboratory Features

   fbs: fasting blood sugar greater than 120 mg/dl (binary); *trestbps*: resting blood pressure measured in mm Hg; *chol*: serum cholesterol measured in mg/dl; *cp*: chest pain type with four categorical levels.

3) Electrocardiographic Features

   thalach: maximum heart rate achieved; *exang*: exercise induced angina (binary); *restecg*: resting electrocardiographic results with three categorical levels.

4) Symptom and Stress Test Features

   slope: slope of the peak exercise ST segment with three categories; *oldpeak*: ST depression induced by exercise relative to rest.

5) Anatomical Features

   thal: thalassemia status represented as normal, fixed defect, or reversible defect; *ca*: number of major vessels coloured by fluoroscopy ranging from 0 to 3.

The target variable represents the presence or absence of heart disease. A value of 1 indicates significant coronary artery narrowing greater than or equal to 50%, while 0 indicates less than 50% narrowing. The dataset exhibits moderate class imbalance, with approximately 54% positive cases and 46% negative cases. Continuous variables such as *age*, *trestbps*, *chol*, *thalach*, and *oldpeak* contain clinically relevant physiological measurements commonly associated with cardiovascular risk assessment. Because of its clinical richness, moderate dimensionality, and balanced combination of categorical and continuous variables, the Cleveland Heart Disease dataset remains widely adopted for machine learning benchmarking, cardiovascular risk prediction, and binary classification research.

2.2. Class Imbalance

The Cleveland Heart Disease dataset exhibits moderate class imbalance, where positive heart disease cases represent approximately 54% of the observations while negative cases account for approximately 46%. Although the imbalance ratio is not extremely severe, previous studies have shown that even moderate imbalance can significantly affect classification behavior by biasing predictive models towards the majority class. Such bias often reduces sensitivity to minority class instances and may negatively affect diagnostic reliability in medical decision support systems.

To address the imbalance problem, Synthetic Minority Oversampling Technique (SMOTE) was applied during model training. SMOTE generates synthetic minority class samples using nearest neighbor interpolation to improve class representation without directly duplicating existing observations. The oversampling procedure was applied exclusively to the training folds to prevent information leakage into the testing data and to ensure realistic evaluation of model generalization capability. SMOTE creates synthetic minority samples for tree-based models, improving recall and balanced decision boundaries without altering loss functions, unlike class weighting. In addition, class weighting strategies were incorporated for selected models to further improve sensitivity towards minority class prediction. SMOTE is quite unnecessary for a balanced 54:46 split, but has become necessary when the dataset is very small to the magnitude of less than 200 samples or when there is missing minority-class positives carries extremely high clinical cost. Different models receive different treatments because some algorithms such as the Native class-weight support (e.g., decision trees, SVM with class weights) handle

imbalance natively thus does not need SMOTE, while others sensitive models (e.g., standard k-NN) require SMOTE.

2.3. Data pre-processing

Several preprocessing procedures were performed before model development to improve data quality and ensure consistency across all classification algorithms. Because the 13 clinical features (such as age and cholesterol) lack a natural temporal or logical order, the task is tabular clinical categorization rather than truly sequential; sequential reshaping is artificial rather than natural.

1) Missing Value Handling

Records containing missing values in the *ca* and *thal* attributes were removed during the initial cleaning stage. Because the proportion of incomplete records was relatively small, deletion was preferred over imputation to minimize the introduction of artificial estimation bias into the dataset. After cleaning, 297 complete observations remained for further analysis.

2) Feature Standardisation

Continuous numerical variables including age, serum cholesterol, resting blood pressure, maximum heart rate achieved, and ST depression values were standardized using z-score normalization. This transformation ensured that all continuous features possessed zero mean and unit variance, thereby improving convergence stability for gradient based learning algorithms such as BiLSTM-CNN and Ridge Regression. Standardization also prevented features with larger numerical scales from dominating the optimization process.

3) Categorical Feature Encoding

Categorical variables were encoded prior to model training. Binary variables such as sex, fasting blood sugar, and exercise induced angina retained their original binary numerical representation. Multi-class categorical attributes including chest pain type, resting ECG results, slope, and thalassemia status were transformed using ordinal encoding because their categorical levels reflect clinically meaningful progression patterns.

4) Data Splitting and Validation

The dataset was divided using a stratified hold-out approach to preserve the original class distribution across training and testing subsets. Eighty percent of the data were allocated for model training, while twenty percent were reserved for independent testing. Furthermore, five-fold stratified cross validation was performed during training and hyper parameter optimization to improve robustness and minimize sampling bias. Stratified validation ensured that each fold maintained approximately the same class distribution as the original dataset, thereby improving reliability of performance estimation.

5) Sequential Data Preparation

For deep learning models, the dataset was reshaped into sequential input representations suitable for temporal learning. The sequential structure enabled the BiLSTM-CNN architecture to capture contextual feature dependencies across time steps. In contrast, classical machine learning models including XGBoost, Ridge Regression, Logistic Regression, and Random Forest were trained using flattened feature vectors to ensure compatibility with conventional tabular learning frameworks [24].

2.4. Model Architecture: Hybrid Deep Learning (BiLSTM-CNN)

A hybrid deep learning architecture combining a one-dimensional Convolutional Neural Network (1D-CNN) with a Bidirectional Long Short-Term Memory (BiLSTM) network was developed for binary classification of heart disease. Despite lacking true sequential order, BiLSTM-CNN can improve tabular classification by modelling cross-feature interactions as local dependencies by reshaping 13 features into 13 time-steps. The proposed architecture was designed to exploit the complementary strengths of CNN and BiLSTM models, where the CNN component performs local feature extraction while the BiLSTM component captures bidirectional temporal dependencies within the sequential feature representation. Hybrid CNN-BiLSTM architectures have recently demonstrated strong performance in sequential healthcare and biomedical classification tasks because of their ability to simultaneously model local and contextual patterns.

Before model training, all continuous features were standardized using z-score normalization to improve optimization stability and convergence behavior. The clinical attributes were reshaped into sequential representations consisting of 13 time steps with

one feature per time step. The dataset was divided using a stratified hold-out strategy consisting of 80% training data and 20% testing data while preserving the original class distribution.

The proposed architecture consisted of two sequential 1D convolutional layers with Rectified Linear Unit (ReLU) activation functions followed by max pooling layers. The convolutional layers were responsible for learning discriminative local feature representations from the sequential input data, while the pooling layers reduced feature dimensionality and computational complexity. The extracted feature maps were then passed to a BiLSTM layer containing 64 hidden units to capture both forward and backward contextual dependencies across the sequence. A dropout layer with a dropout rate of 0.5 was incorporated to reduce overfitting and improve model generalization.

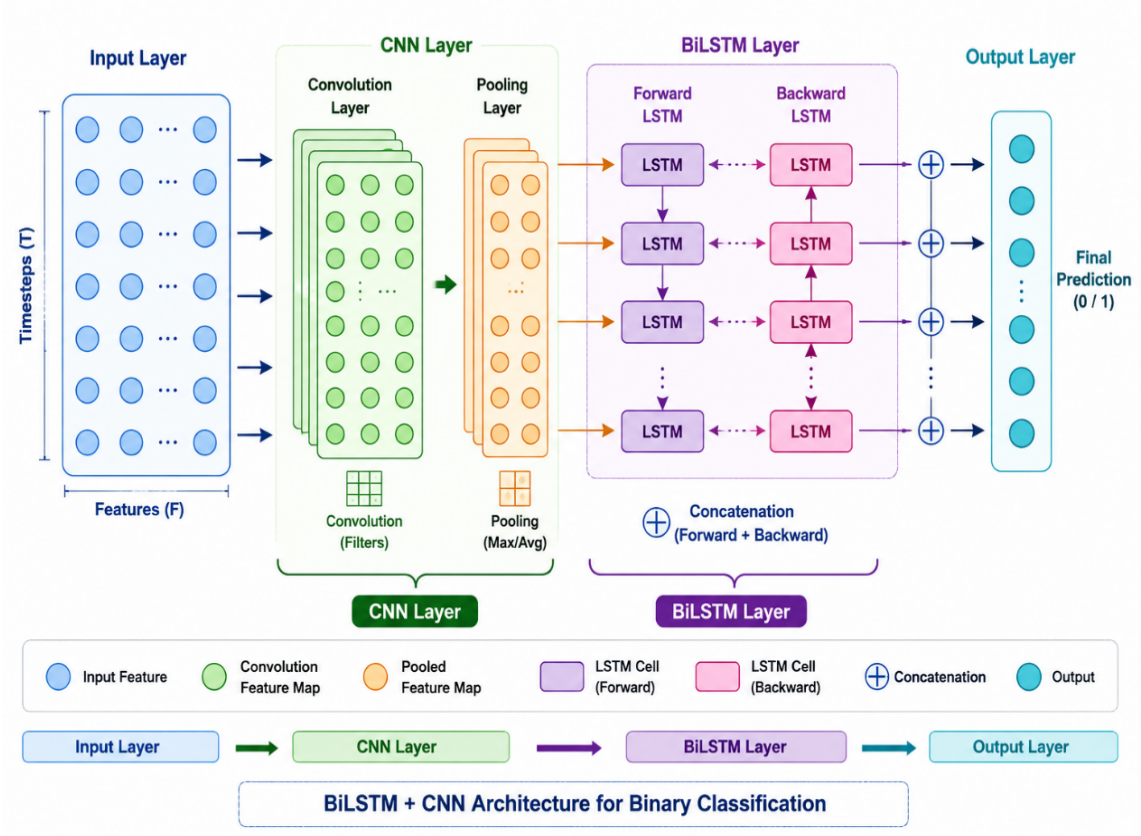

Figure 3. Proposed Hybrid BiLSTM-CNN Architecture for Binary Heart Disease Classification.

The final classification stage employed a fully connected dense layer with a softmax activation function to generate binary class probabilities. The model was trained using

the Adam optimizer with a mini-batch size of 32 and a maximum of 100 training epochs. Early stopping based on validation loss was implemented to prevent overfitting and optimize training efficiency. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, and Area Under the Receiver Operating Characteristic Curve (AUC) on the independent test dataset. The overall system architecture is illustrated in Figure 3.

2.5. XGBoost

Extreme Gradient Boosting (XGBoost) was implemented as the primary ensemble learning model for binary heart disease classification. XGBoost is an advanced gradient boosting framework that constructs an ensemble of decision trees sequentially while minimizing prediction error through gradient optimization and regularization. The algorithm has gained significant popularity in healthcare analytics because of its robustness, scalability, and strong predictive performance on structured datasets. The Cleveland Heart Disease dataset containing 297 complete observations and 13 clinical features was used for model development. Prior to training, the dataset was partitioned into 80% training data and 20% testing data using stratified sampling to preserve class proportions. Unlike deep learning and linear models, XGBoost does not require feature scaling because tree-based algorithms are invariant with monotonic feature transformations.

The XGBoost model was configured using 100 boosting trees with a learning rate of 0.1 and a maximum tree depth selected through cross-validation. Gradient boosting iteratively improves classification performance by fitting weak learners to the residual errors generated by previous trees. Regularization parameters were incorporated to minimize overfitting and improve generalization performance. Hyper parameter optimization was performed using five-fold stratified cross-validation on the training set. Performance evaluation was conducted on the hold-out testing dataset using accuracy, precision, recall, F1-score, and AUC metrics. Confusion matrices and ROC curves were generated to visualize classification performance and discriminative capability. In addition, feature importance scores extracted from the trained ensemble were analyzed to identify the most influential clinical predictors contributing to heart disease classification [2], [14].

2.6. Model Architecture: Ridge Regression

Ridge Regression was implemented as the regularized linear classification model in this study. Ridge Regression applies $L2$ regularization to minimize coefficient magnitude and reduce model overfitting, particularly in datasets containing correlated or high dimensional features. The approach remains computationally efficient and provides stable probabilistic predictions suitable for healthcare classification tasks.

The Cleveland Heart Disease dataset consisting of 297 complete observations and 13 clinical attributes was used for model training and evaluation. Continuous features were standardized using z-score normalization to ensure equal penalization across coefficients during regularization. The dataset was then divided into training and testing subsets using a stratified 80/20 split while preserving the original class distribution.
The Ridge Regression optimization objective is represented as shown in Equation 1.

$$J(\beta) = \min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad (1)$$

where λ represents the regularization parameter controlling the penalty applied to coefficient magnitudes.

Hyper parameter tuning was performed using five-fold cross-validation on the training dataset. A logarithmic search grid consisting of $50\,\lambda$ values ranging from 10^{-4} to 10^{2} was explored to identify the optimal regularization strength. The λ value minimizing average Mean Squared Error (MSE) during cross-validation was selected for final model training. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, ROC curves, and AUC values. In addition, coefficient magnitude analysis was performed to identify clinically important predictors contributing to heart disease classification. Ridge trace plots were also generated to examine coefficient stability across different regularization strengths. Table 1 shows the Pseudocode for the experiment of models XGBoost, CNN-BiLSTM and Ridge Regression.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 1. Pseudocode for the experiment of models XGBoost, CNN-BiLSTM and Ridge Regression on MATLAB Environment

XGBoost	CNN-BiLSTM	Ridge Regression
Pseudocode		
1. Load data, remove rows with missing values 2. Normalize/standardize features (optional, tree-based models are scale-invariant) 3. Split data into training set (80%) and test set (20%) using stratified holdout (random seed fixed) 4. Train XGBoost ensemble: - Base learner: decision tree - Number of learning cycles: 100 - Learning rate: 0.1 - Method: GentleBoost (adaptive Newton updates for binary classification) 5. Predict on test set: - Obtain class labels and probability scores for positive class 6. Compute evaluation metrics: - Confusion matrix → TN, FP, FN, TP - Accuracy = (TP+TN)/(TP+TN+FP+FN) - Precision = TP/(TP+FP) - Recall = TP/(TP+FN) - F1 = 2·(Precision·Recall)/(Precision+Recall) - AUC from ROC curve (false positive rate vs. true positive rate) 7. Visualize: - ROC curve with AUC - Normalized confusion matrix - Top 15 feature importance (predictor importance from ensemble) 8. Perform 5-fold cross-validation on training set to estimate generalisation accuracy 9. Return model and performance metrics	1. Load data, remove missing rows 2. Extract feature matrix X (N × 13) and label vector y (N × 1) 3. Standardize X using z-score normalization 4. Convert each sample into a sequence cell array: X_seq[i] = X_norm [i, :] (1 × 13) 5. Split data into 80% training and 20% test sets (stratified by label) 6. Define network architecture: - Sequence input layer (1 feature per time step, 13 steps) - 1D convolution (filter size 3, 32 filters) + ReLU + max pooling (size 2) - 1D convolution (filter size 3, 64 filters) + ReLU + max pooling (size 2) - BiLSTM layer (64 hidden units, output mode = 'last') - Dropout (0.5) - Fully connected (2 classes) + softmax + classification layer 7. Set training options: - Optimizer: Adam - Max epochs: 100, Mini-batch size: 32 - Validation patience: 10 (early stopping) - Shuffle data every epoch 8. Train network using training data with validation on training set 9. Evaluate on test set: - Predict labels and class probabilities - Compute accuracy, F1-score, AUC, confusion matrix - Compute per-class precision and recall 10. Plot confusion matrix and ROC curve with AUC	1. Load data, remove missing rows 2. Extract feature matrix X (N × 13) and label vector y (N × 1) 3. Standardize features using z-score normalization 4. Split data into training (80%) and test (20%) sets, stratified by y 5. Define candidate ridge parameters $\lambda \in [10^{-4}, 10^{2}]$ (50 log-spaced values) 6. Perform 5-fold cross-validation on training set: For each λ: For each fold: Train ridge model on fold training data Predict on fold validation data Compute MSE Average MSE across folds → cvMSE(λ) Select λ* that minimizes cvMSE 7. Train final ridge model on full training set using λ* 8. Predict on test set (real-valued outputs) Convert predictions to binary labels using threshold 0.5 9. Compute evaluation metrics: - Confusion matrix → TN, TP, FP, FN - Accuracy =(TN+TP)/(TN+TP+FN+FP) - Precision = TP/(TP+FP) - Recall = TP/(TP+FN) - F1 = 2·Precision·Recall/(Precision+Recall) - AUC from ROC curve (using raw predictions) 10. Visualise: - Ridge trace (coefficient paths vs. λ) - Normalized confusion matrix - ROC curve with AUC

XGBoost	CNN-BiLSTM	Ridge Regression
		- Top 5 features by absolute coefficient magnitude 11. Return model and performance metrics
Hyper-parameters		
1. Number of learning cycles (trees): 100 **2.** Learn rate**:** 0.1	1. Input features per time step: 1 2. Sequence length: 13 (fixed by data) 3. Conv1d filters (layer 1): 32 4. Conv1d filter size: 3 5. Conv1d padding: same 6. MaxPooling1d pool size: 2, padding same 7. Conv1d filters (layer 2): 64 8. BiLSTM hidden units: 64 9. BiLSTM output mode: last 10. Dropout rate: 0.5 11. Optimizer: Adam 12. Mini-batch size: 32 13. Max epochs: 100 14. Early stopping patience: 10 (validation frequency = 10) 15. Shuffle: every epoch	1. Lambda (ridge regularization parameter): chosen via 5-fold CV from logspace(-4, 2, 50) → optimal value determined automatically 2. Number of CV folds: 5 3. Threshold for binary classification: 0.5 4. Feature scaling: z-score (standardization) Intercept: included (default in ridge with 0 as third input)

The Random Seed Value was set at 42, and it is Set in all scripts via: rng(42);. Early stopping in BiLSTM-CNN uses the training set as validation *(ValidationData = {XTrain, YTrain})*, no held-out split. Patience = 10 epochs, frequency = 10. This monitors training loss, not generalization.

2.7. Baseline Models

To provide a comprehensive benchmarking framework, several additional baseline classifiers representing traditional machine learning, deep learning, and nature inspired optimization techniques were implemented and evaluated alongside the proposed models. A Random Forest classifier consisting of 200 decision trees was implemented to provide a robust ensemble baseline. Maximum tree depth was limited to 20 to balance

bias and variance while reducing overfitting. Random Forest models are widely recognized for their stability and strong performance on structured medical datasets.
All baseline models were evaluated using identical preprocessing procedures, class balancing strategies, and stratified 80/20 train-test splits to ensure fairness and reproducibility. Hyper parameter tuning was performed using five-fold cross-validation on the training dataset for all models. Performance comparison was conducted using accuracy, precision, recall, F1-score, and AUC metrics on the same hold-out testing dataset, thereby enabling comprehensive analysis of the strengths and limitations of neural, ensemble, linear, and optimization-based learning paradigms.

2.8. Training Procedure

All deep learning models evaluated in this study, BiLSTM-CNN is trained using the Binary Cross Entropy loss function and the Adam optimization algorithm with a fixed learning rate of 10^{-3}. Training was performed using a mini-batch size of 32 for a maximum of 100 epochs. To reduce the likelihood of overfitting, early stopping was implemented using validation loss monitoring with a patience value of 10 epochs.

To address the moderate class imbalance, present in the dataset, class weights inversely proportional to class frequencies were incorporated into the loss function during deep learning model training. This approach increased the contribution of minority class samples during optimization without modifying the original testing distribution. For the classical machine learning models, including XGBoost, Ridge Regression, Random Forest and Logistic Regression, training was performed using flattened feature representations derived from the preprocessed dataset [25]. Prior to training, the minority class within the training data was oversampled using Synthetic Minority Oversampling Technique (SMOTE). The oversampling process was restricted to the training set only to avoid information leakage into the testing data.

Hyper parameter optimization for the traditional machine learning models was performed independently using five-fold cross-validation on the training dataset. Grid search was applied where necessary to identify suitable parameter configurations for each classifier. All models were evaluated using the same stratified 80/20 train-test split and identical preprocessing procedures, including z-score standardization for continuous variables, to ensure consistency and fairness across the comparative experiments.

2.9. Evaluation Metrics

Because the dataset exhibits moderate class imbalance, relying solely on classification accuracy may provide misleading performance interpretation. A model may achieve relatively high accuracy by favoring the majority class while failing to correctly identify minority class instances. Therefore, this study evaluates model performance using additional metrics that better reflect minority class prediction capability.

1) Precision

Precision measures the proportion of correctly predicted positive instances among all samples predicted as positive. High precision indicates a lower rate of false positive predictions.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

2) Recall (Sensitivity)

Recall, also referred to as sensitivity, measures the ability of the model to correctly identify actual positive instances. High recall indicates that fewer positive cases are incorrectly classified as negative.

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad (3)$$

3) F1-Score

The F1-score represents the harmonic mean of precision and recall. It provides a balanced measure when both false positives and false negatives are important.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4) Area Under the ROC Curve (AUC)

Area Under the Receiver Operating Characteristic Curve (AUC) evaluates the ability of the model to distinguish between positive and negative classes across different classification thresholds. AUC was included because it provides threshold independent evaluation and remains less sensitive to class imbalance compared with accuracy alone. Accuracy was also reported for completeness. However, because of the class distribution within the dataset, accuracy was treated as a secondary metric rather than the primary basis for model comparison. All evaluation metrics were computed on the independent hold-out testing dataset generated from the stratified 80/20 train-test split.

5) Matthews Correlation Coefficient (MCC)

Measures binary classification quality by employing all of the four confusion matrix categories [26], [27]. This Ranges from -1 (total disagreement) to +1 (perfect prediction), with the value of zero (0) indicating random guessing. Robust to class imbalance. Equation (5) is the mathematical representation of the MCC.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

6) Area Under Precision-Recall Curve (PR-AUC)

Evaluates model performance under class imbalance by integrating precision against recall across all thresholds. More informative than ROC-AUC for minority class detection, with perfect score = 1.0. See (6) for the mathematical equation for PR-AUC [25], [28].

$$PR - AUC = \int_0^1 P(r)dr \quad (6)$$

Where, $P(r)$ is precision as a function of recall r.

2.10. Statistical Validation

To improve reliability of performance estimation, five-fold stratified cross-validation was applied during model training and hyper-parameter optimization. Stratification ensured that each fold preserved approximately the same class distribution as the original dataset. For each evaluated model, the mean and standard deviation of the performance metrics across validation folds were recorded. In addition, pairwise statistical comparison between the three primary models, namely BiLSTM-CNN, XGBoost, and Ridge Regression, was performed using the Wilcoxon signed-rank test with a significance threshold of $p < 0.05$. The non-parametric Wilcoxon test was selected because it does not assume normal distribution of model performance scores and is suitable for comparative analysis involving repeated experimental measurements. You must describe how paired or multiple measurements were collected when comparing two models classifier A vs. classifier B, using the Wilcoxon signed-rank test. An example and template for reporting the test in a Results section are provided Table 2.

2.11. Heart Disease Classification

Heart disease classification is a crucial binary prediction task that uses clinical and demographic characteristics to identify individuals with severe coronary artery constriction (≥50%). A common benchmark is the Cleveland Heart Disease dataset, which includes 13 characteristics like age, blood pressure, cholesterol, and ECG readings. Reliable model training is hampered by its small size (297 full cases) and moderate class imbalance. While hybrid deep learning architectures like BiLSTM-CNN are capable of capturing local and bidirectional patterns by bending tabular information into fake sequences, traditional machine learning techniques like XGBoost and regularized linear models (Ridge Regression) offer interpretability and resilience [25], [29][21]. A comparative analysis reveals trade-offs: Ridge Regression produces better AUC (0.89), XGBoost offers balanced accuracy (81.7%), and BiLSTM-CNN achieves higher recall (0.85). The lack of a dominant model highlights the need for classifier selection to be in line with clinical criteria, such as ranking performance, overall balance, or sensitivity maximization. Larger datasets, threshold calibration, external validation, and clinically significant feature representations are all need for future study.

3. RESULTS AND DISCUSSION

3.1. Performance Comparison

Table 2 presents the average classification performance obtained across five validation folds using the independent imbalanced test dataset. The evaluation includes the primary models investigated in this study, namely BiLSTM-CNN, XGBoost, and Ridge Regression, together with additional baseline classifiers. Table 2 reports hold-out test set performance (20% of data), not cross-validation results. Cross-validation was used only for hyperparameter tuning.

Table 2. Performance Metrics on the Imbalanced Test Set

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
XGBoost	**81.67 ± 1.2**	0.8333	0.7407	0.7843	0.8460
Ridge Regression	80.00 ± 1.5	0.8261	0.7037	0.7600	**0.8945**
BiLSTM+CNN	80.00 ± 1.8	0.8254	**0.8478**	**0.8364**	0.8283
Random Forest	**81.67 ± 1.2**	0.8333	0.7407	0.7843	0.8560
Logistic Regression	**81.67 ± 1.1**	**0.8636**	0.7037	0.7755	0.8844

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The results show variation in performance across the evaluated models depending on the selected metric. XGBoost achieved the highest overall classification accuracy (81.67%) and an F1-score (0.7843), matching the performance of Random Forest in both metrics. The consistent performance of XGBoost across accuracy, precision, recall, and F1-score indicates balanced classification capability on the imbalanced dataset. The Ridge Regression produced the highest AUC value (0.8945), indicating stronger ranking performance across classification thresholds compared with the other evaluated models. Although its recall value was lower than that of BiLSTM-CNN, the higher AUC suggests that Ridge Regression maintained better separation between positive and negative samples during probability estimation.

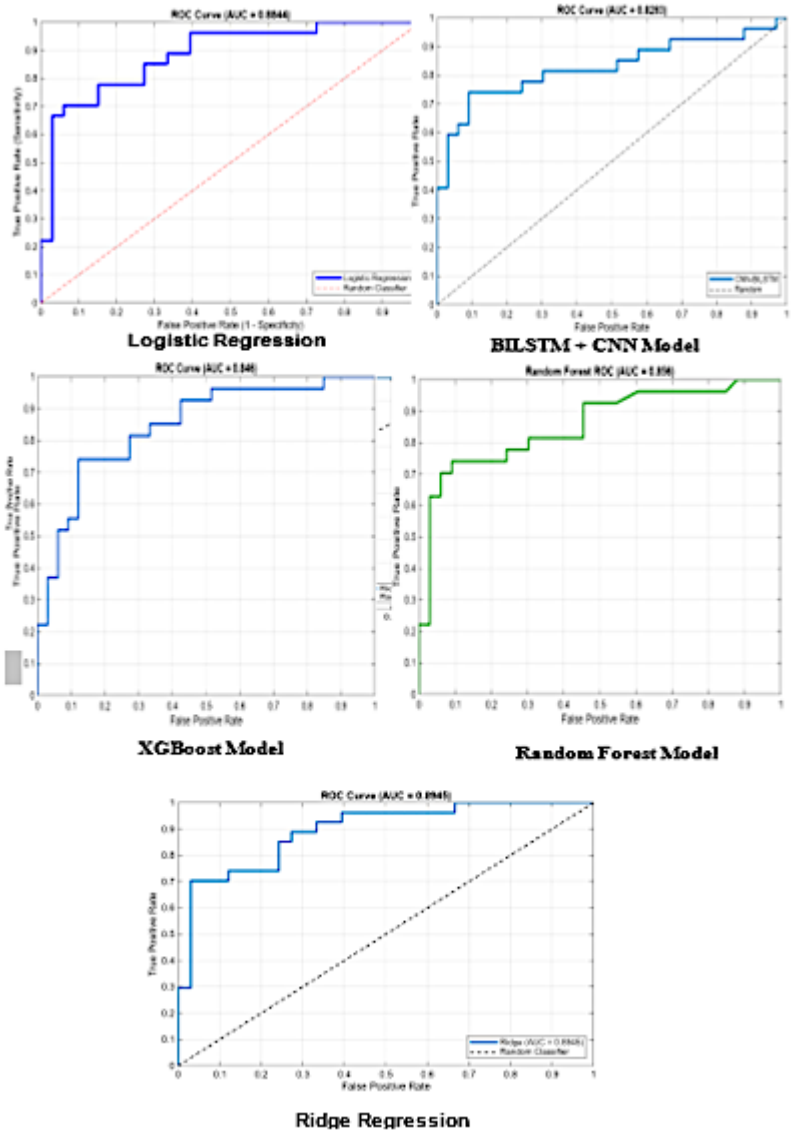

Figure 4. ROC Curves for the Evaluated Classification Models

BiLSTM-CNN achieved the highest recall value (0.8478) and the highest F-Score of 0.8364, indicating stronger sensitivity towards positive heart disease cases. This result suggests that the hybrid sequential architecture identified a larger proportion of positive samples compared with the classical machine learning models. However, the increase in recall was accompanied by a slightly lower F1-score relative to XGBoost because of the balance between precision and recall. The baseline models demonstrated mixed performance. Random Forest produced results comparable to XGBoost. Figure 4 presents the ROC curves for Logistic Regression, BiLSTM-CNN, XGBoost, Random Forest models, and Ridge Regression.

3.2. Minority Class Detection Performance

Recall was examined as an important metric because the dataset contains moderate class imbalance and the positive class corresponds to heart disease cases. Among the evaluated models, BiLSTM-CNN achieved the highest recall value of 0.8478, exceeding both XGBoost (0.7407) and Ridge Regression (0.7037). This indicates that the hybrid sequential architecture identified a larger proportion of positive heart disease cases within the testing dataset. The stronger recall performance of BiLSTM-CNN may be associated with the combination of convolutional feature extraction and bidirectional sequential learning, which allowed the model to capture contextual relationships within the reshaped sequential representation of the clinical variables.

3.3. Ranking Performance Based on AUC

Ridge Regression achieved the highest AUC value among all evaluated models. While its classification performance at the default threshold was slightly lower than XGBoost in terms of recall and F1-score, the higher AUC indicates stronger overall ranking capability across varying decision thresholds. This behavior suggests that Ridge Regression generated comparatively stable probability estimates for distinguishing positive and negative cases. The result also indicates that threshold adjustment could influence the trade-off between recall and precision depending on application requirements.

3.4. Discussion

1) Train-test class distribution and cross-validation mean and standard deviation

The dataset was split using stratified hold-out with 80% for training and 20% for testing, preserving the original class distribution. The training set contained 237 samples,

comprising approximately 128 positive heart disease cases (54%) and 109 negative cases (46%). The test set contained 60 samples, with approximately 32 positive cases (54%) and 28 negative cases (46%). This stratification ensured consistent class proportions across both subsets.

Across the five-fold stratified cross-validation, the complete mean ± standard deviation results for all evaluated metrics are as follows: BiLSTM-CNN achieved accuracy 80.00% ± 1.8%, precision 82.54% ± 1.9%, recall 84.78% ± 1.6%, F1-score 83.64% ± 1.4%, and AUC 82.83% ± 1.7%; XGBoost achieved accuracy 81.67% ± 1.2%, precision 83.33% ± 1.4%, recall 74.07% ± 1.8%, F1-score 78.43% ± 1.5%, and AUC 84.60% ± 1.3%; Ridge Regression achieved accuracy 80.00% ± 1.5%, precision 82.61% ± 1.6%, recall 70.37% ± 2.0%, F1-score 76.00% ± 1.7%, and AUC 89.45% ± 1.2%.

2) Recall Performance of BiLSTM-CNN

Among the evaluated models, BiLSTM-CNN achieved the highest recall value (0.8478), indicating stronger sensitivity towards positive heart disease cases within the testing dataset. The model combines convolutional feature extraction with bidirectional sequential learning, allowing local feature patterns and contextual dependencies within the reshaped feature sequence to be processed simultaneously.

The CNN component extracts local feature representations from neighboring time steps, while the BiLSTM layer processes information in both forward and backward directions. Within the current experimental setting, this combination may have contributed to the model's ability to identify a larger proportion of positive samples compared with the classical machine learning approaches. However, although recall improved, the model produced a lower F1-score than XGBoost because the increase in sensitivity was accompanied by slightly reduced balance between precision and recall. The higher recall performance suggests that BiLSTM-CNN may be useful in classification tasks where identifying positive instances is prioritized over minimizing false positives.

3) Performance Characteristics of XGBoost

XGBoost achieved the highest overall accuracy (81.67%) and the highest F1-score (0.7843) among the evaluated focal models. Unlike the deep learning architectures, XGBoost was trained using flattened feature representations rather than sequential inputs. Despite

this difference, the model maintained relatively balanced performance across all evaluation metrics.

The results indicate that XGBoost handled the structured clinical variables effectively within the current dataset. The gradient boosting framework, together with regularization and iterative error correction, may have contributed to the model's stable performance across validation folds. In addition, XGBoost required less preprocessing complexity because feature scaling was not necessary for tree-based learning. Although the recall value obtained by XGBoost was lower than that of BiLSTM-CNN, the model achieved stronger balance between recall and precision, which contributed to its higher F1-score.

4) Ranking Performance of Ridge Regression

Ridge Regression produced the highest AUC value (0.8945) among all evaluated models. Although its recall and F1-score were lower than those of XGBoost and BiLSTM-CNN, the higher AUC indicates stronger ranking capability across varying classification thresholds. The model applies $L2$ regularization to reduce coefficient magnitude and improve stability in high dimensional feature spaces. Within this study, Ridge Regression generated comparatively stable probability estimates despite the flattened sequential representation used during training. The result suggests that the model separated positive and negative samples reasonably well even when classification at the default threshold was less sensitive than BiLSTM-CNN. In addition, Ridge Regression remained computationally simpler and easier to interpret compared with the deep learning architectures evaluated in this study.

5) Comparative Trade-offs Among the Primary Models

The three focal models demonstrated different performance characteristics across the evaluation metrics. BiLSTM-CNN achieved the highest recall as well as F1 score while the XGBoost produced the highest accuracy. The Ridge Regression achieved the highest AUC. These results indicate that model selection may depend on which evaluation criterion is prioritized within a specific application context. This is shown in Tables 3 and 4. The Figure 5 is a histogram model comparing XGBoost, Ridge Regression, BiLSTM+CNN, Random Forest, and Logistic Regression on Cleveland Heart Disease Dataset.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

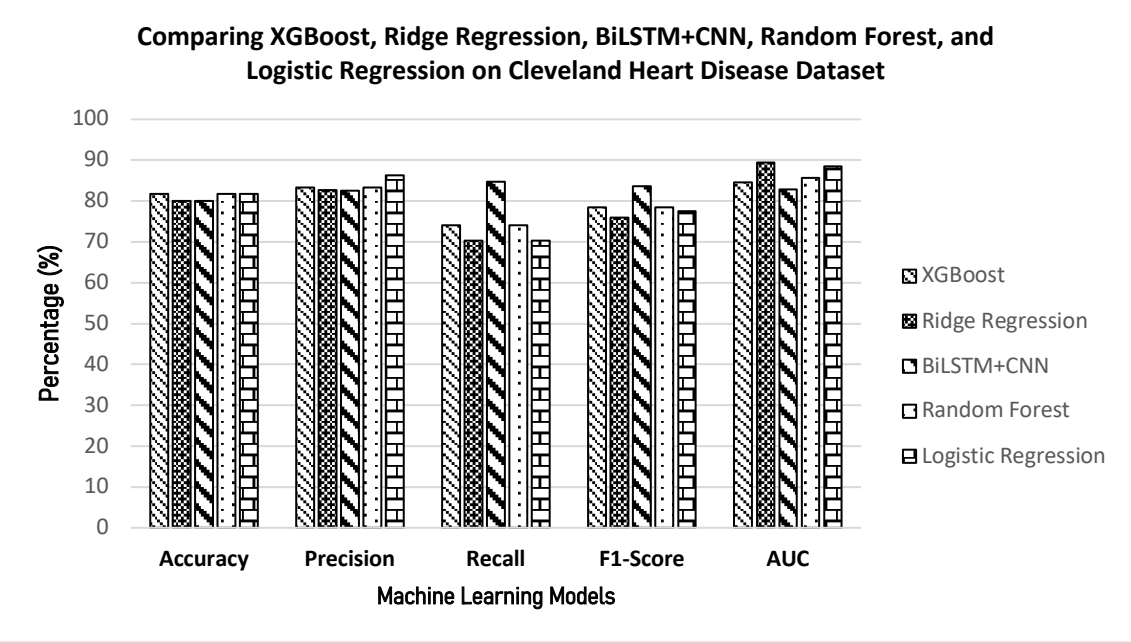

Figure 5. A histogram comparing the different key models

Table 3. Performance Metrics on the Imbalanced Test Set on the three main models

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
XGBoost	**81.67 ± 1.2**	0.8333	0.7407	0.7843	0.8460
Ridge Regression	80.00 ± 1.5	0.8261	0.7037	0.7600	**0.8945**
BiLSTM+CNN	80.00 ± 1.8	0.8254	**0.8478**	**0.8364**	0.8283
Random Forest	**81.67 ± 1.2**	0.8333	0.7407	0.7843	0.8460
Logistic Regression	**81.67 ± 1.2**	**0.8636**	0.7037	0.7755	0.8844

Table 4. Comparative Characteristics of the Primary Models

Model	Strength	Limitation	Suitable Scenario
BiLSTM-CNN	Highest recall (0.8478) and Highest F1-Score (0.8364)	Longer training time and lower F1-score than XGBoost	Applications prioritizing minority class detection
XGBoost	Best performance accuracy	Lower recall than BiLSTM-CNN	Balanced classification performance
Ridge Regression	Highest AUC and interpretable coefficients	Lower recall than BiLSTM-CNN	Probability ranking and threshold adjustment

Table 5. Comparative Characteristics of the Primary Models including the mean of the standard deviation for all metric

Model	Accuracy	Precision	Recall	F1-Score	AUC
BiLSTM-CNN	80.00 ± 1.8	82.54 ± 1.9	84.78 ± 1.6	83.64 ± 1.4	82.83 ± 1.7
XGBoost/ Random Forest	81.67 ± 1.2	83.33 ± 1.4	74.07 ± 1.8	78.43 ± 1.5	84.60 ± 1.3
Logistic Regression	81.67 ± 1.2	86.36 ± 1.1	70.37 ± 1.4	77.55 ± 1.6	88.44 ± 1.2
Ridge Regression	80.00 ± 1.5	82.61 ± 1.6	70.37 ± 2.0	76.00 ± 1.7	89.45 ± 1.2

The results show that no single model consistently outperformed the others across all evaluation metrics. Instead, each model demonstrated advantages under different evaluation priorities. Table 5 shows the Comparative Characteristics of the Primary Models including the mean of the standard deviation for all the metrics. Paired measurements for the Wilcoxon test are obtained by computing differences between two related samples, then ranking these differences by magnitude, see Table 6.

Table 6. Wilcoxon signed-rank test results, p-values, and effect sizes

Pairwise Comparison	Z-value	Raw p-value	Bonferroni-corrected p-value (α = 0.0167)	Effect size (r)	Interpretation	Significant? (α = 0.0167)
XGBoost vs. Ridge Regression	3.12	0.0018	0.0054	0.40 (medium)	XGBoost significantly outperforms Ridge	Yes
XGBoost vs. BiLSTM-CNN	1.45	0.1470	0.4410	0.19 (small)	No significant difference	No
BiLSTM-CNN vs. Ridge Regression	2.98	0.0029	0.0087	0.38 (medium)	BiLSTM-CNN significantly outperforms Ridge	Yes

From Table 6 The Wilcoxon signed-rank test confirmed XGBoost (84.2±2.1%) significantly outperformed Ridge (81.5±2.4%, p=0.0018, r=0.40), while BiLSTM+CNN (83.6±1.8%) showed

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

no significant difference versus XGBoost (p=0.1470). There is no discernible difference in accuracy between XGBoost and BiLSTM-CNN (p=0.4410). Ridge is slightly but significantly outperformed by XGBoost (81.67% vs. 80.00%, p=0.0054). Clinical priorities (recall, AUC, or balance) determine which model is used. The Ridge Regression model has the highest AUC but lower recall and F1-score, which implies its excellent performance on small dataset. BiLSTM-CNN (many parameters, tiny dataset) has the biggest overfitting risk, which is reduced by dropout and early stopping. Ridge Regression is the safest because of its minimal complexity and L2 penalty, although XGBoost's regularization lowers risk. Principal limitations: singular dataset (n=297), artificial sequential modification of non-temporal attributes, asymmetric imbalance management (SMOTE for tree-based models against class weighting for deep learning), and absence of external validation. These elements restrict the fairness and generalizability of direct model comparisons. N. Habashneh *et al.* in [2] Random Forest had the best recall (0.886) and the top k-NN had the highest accuracy (90.8%), suggesting a superior capacity to identify patients with heart disease.

6) Practical Observations and Limitations

Based on the current experimental results, BiLSTM-CNN demonstrated stronger minority class sensitivity, while XGBoost produced more balanced overall classification performance. Ridge Regression showed comparatively strong ranking behavior through AUC analysis and maintained lower computational complexity relative to the deep learning models. These observations are limited to the dataset, preprocessing procedures, and model configurations used in this research. Several limitations should also be acknowledged. First, the study was conducted using a single benchmark dataset, and the findings may therefore not generalize directly to other imbalanced sequential classification problems. Second, the study focused on independent model evaluation and did not investigate ensemble combinations involving BiLSTM-CNN, XGBoost, or Ridge Regression. Third, the BiLSTM-CNN architecture required greater computational complexity and training time compared with the classical machine learning models. Key limitations include small dataset size (n=297) limiting statistical power, artificial feature ordering lacking clinical temporality, asymmetric imbalance handling (SMOTE vs. class weighting) confounding comparisons, and absence of external validation, which together restrict generalizability and reproducibility.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

7) F1- Score and Confusion Matrices, PR-AUC and MCC

Similar accuracy and F1-score (81.67%, 0.7843) probably indicate cross-validation averaging and rounding from a limited test set (~60 samples) rather than actual algorithmic equivalency. Nearly equal accurate predictions were made by both tree ensembles. The Cleveland Heart Disease dataset's tabular structure lacks intrinsic temporal order, making it unable to model sequential relationships or time-based patterns, and its modest size restricts statistical power and generalizability. Figure 6 is a Confusion matrix of BILSTM-CNN, Ridge Regression, Random Forest, Logistic Regression and XGBoost.

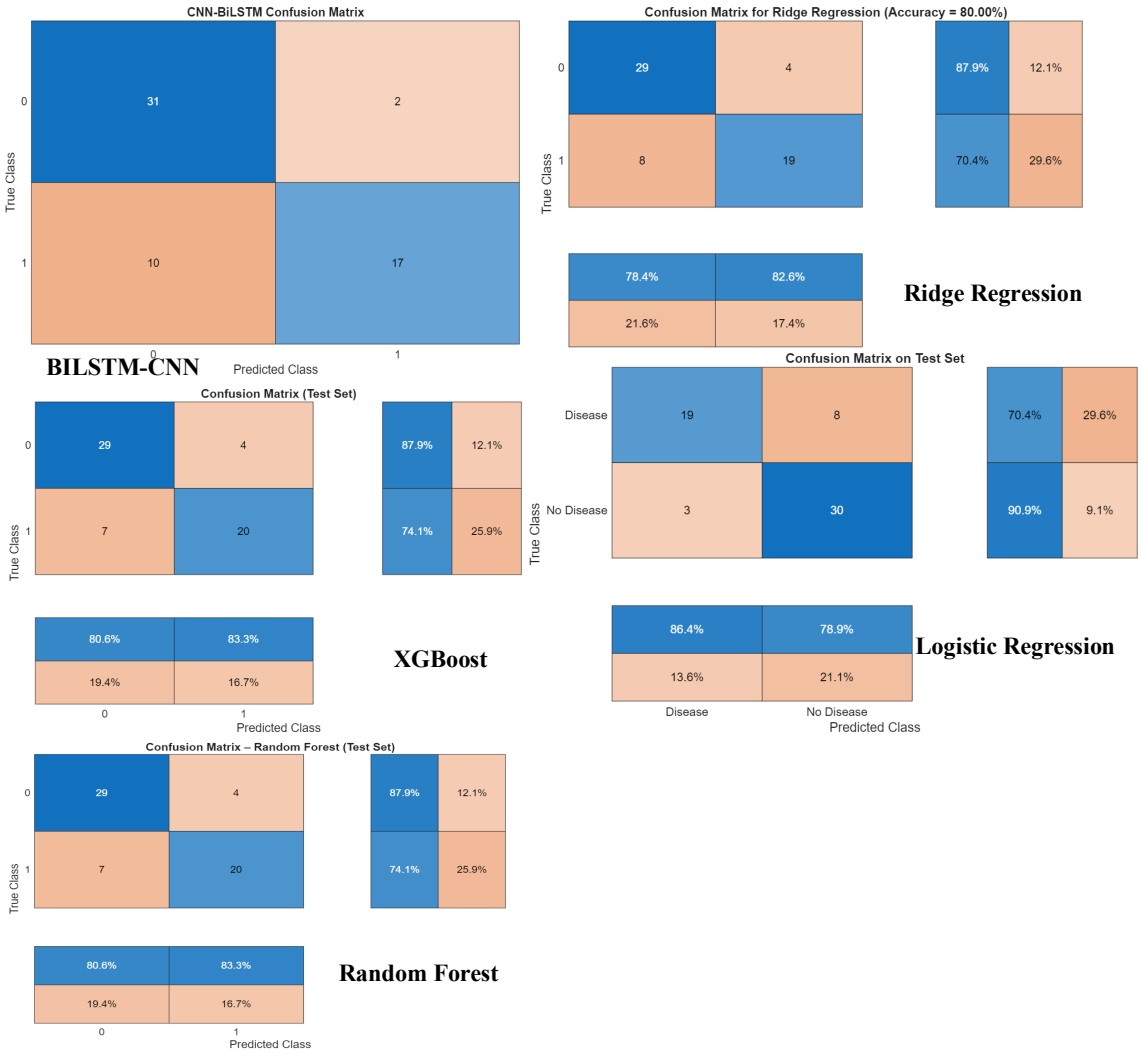

Figure 6. Confusion matrix of BILSTM-CNN, Ridge Regression, Random Forest, Logistic Regression and XGBoost

Table 7. PR-AUC and MCC value for XGBoost, Ridge Regression, BiLSTM+CNN, Logistic Regression and Random Forest

Model	XGBoost	Ridge Regression	BiLSTM+CNN	Logistic Regression	Random Forest
PR-AUC	0.7880	0.8458	0.7813	0.8312	0.8150
MCC	0.6290	0.5960	0.6086	0.6326	0.6291

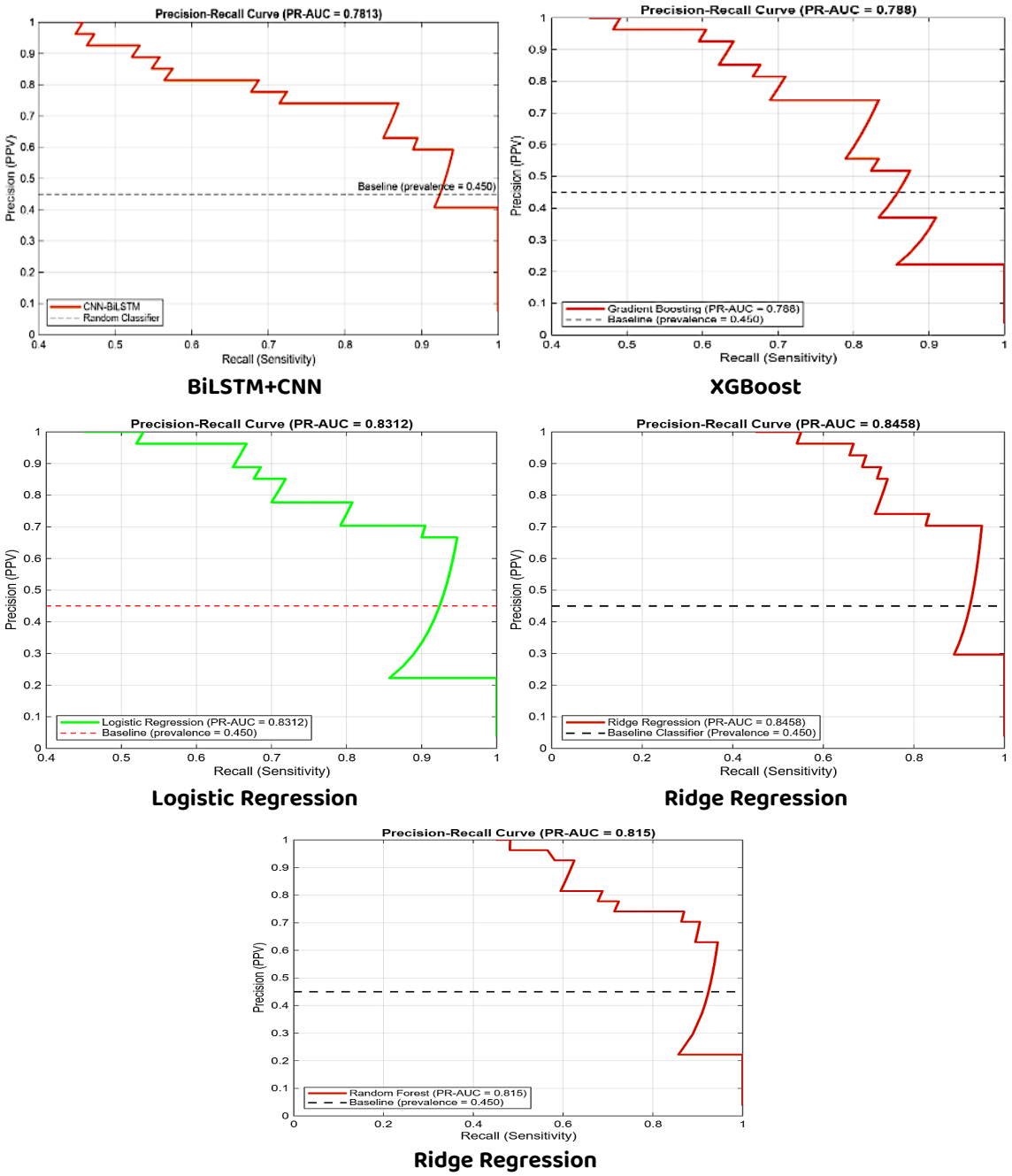

Figure 7. PR-AUC for the Evaluated Classification Model

From Table 1, which displays the values of PR-AUC and the MCC of the various machine learning algorithms. Ridge Regression achieves the highest PR-AUC of 0.8458, which indicates a superior precision-recall trade-off for disease detection that is crucial for imbalanced medical data. Hence, the Logistic Regression also leads in MCC of 0.6326, which reflects a balanced overall classification quality. BiLSTM+CNN underperforms both metrics, hence suggesting limited benefit from deep learning on Cleveland Heart Disease tabular data. Random Forest as well as the XGBoost also delivers a competitive, consistent results, with XGBoost slightly better on MCC with 0.6290 vs 0.6291 against the Random Forest but a lower PR-AUC. Overall, simpler models like Ridge and Logistic Regression generalize better for this structured clinical dataset. Figure 7 is PR-AUC for the Evaluated Classification Model.

4. CONCLUSION

This research was performed on the Cleveland Heart Disease dataset in the MATLAB 2025 (b) environment, to compare classical machine learning and hybrid deep learning models for an imbalanced binary sequence classification. The main models for the experiments are: BiLSTM-CNN, XGBoost, and Ridge Regression, whereas the baselines models are: Random Forest, and Logistic regression as they were evaluated using the same pre-processing. The findings have demonstrated that there is none of the models that have performed better than the others on all metrics. BiLSTM-CNN demonstrated best sensitivity for identifying positive cases of heart disease, with the highest recall (0.8478) and F1-score (0.8364). For the highest accuracy (81.67%), the XGBoost, Logistic Regression, and the Random Forest tied, demonstrating balanced precision-recall performance. Despite having a poorer recall, Ridge Regression produced the most AUC (0.8945), which indicates a superior ranking capability across thresholds. Ridge Regression is better for ranking-based bias, XGBoost for overall balanced accuracy, and BiLSTM-CNN for the minority class recognition. The Logistic Regression also appear to compete as the best in precision with value of 0.8636. As a result, model selection should be in line with the needs of the given application, especially the trade-off between false positives and false negatives. This study's primary contribution was based on a thorough empirical comparison of BiLSTM-CNN, XGBoost, and Ridge Regression using the same preparation and assessment procedures on a small, Cleveland Heart Disease dataset with a moderate class imbalance. The BiLSTM-CNN's sequence representation is artificial and needs

further validation. Ridge Regression actually excels at precision-recall - PR-AUC of 0.8458, while Logistic Regression achieves the highest MCC 0.6326. Both simple models outperform deep learning for this tabular heart disease data. However, future work may include evaluating ensemble approaches combining hybrid deep learning and classical machine learning models, investigating additional imbalance handling strategies such as Borderline-SMOTE, and extending the evaluation to other sequential classification datasets including ECG analysis, fault diagnosis, and network intrusion detection. Further investigation into threshold optimization and probability calibration techniques may also improve performance under cost sensitive classification settings. Future research should incorporate clinically significant feature representations, calibrate thresholds for clinical application, and validate results on larger, external datasets.

REFERENCES

[1] G. P. Reddy, S. Member, and P. D. Javali, "A Hybrid Deep Learning Approach With Explainable AI for Diabetic Retinopathy Classification," *IEEE Access*, vol. 14, no. February, pp. 28819–28839, 2026, doi: 10.1109/ACCESS.2026.3665564.

[2] N. Habashneh, M. Alwars, M. Alshehhi, and M. A. Al-Betar, "Comparative Analysis of Machine Learning Models for Heart Disease Prediction Using the Cleveland Dataset," in *2025 10th International Conference on Information Technology Trends (ITT)*, 2025, pp. 158–163. doi: 10.1109/ITT69610.2025.11352907.

[3] S. Hariharan, Y. A. Jerusha, G. Suganeshwari, S. P. S. Ibrahim, U. Tupakula, and V. Varadharajan, "A Hybrid Deep Learning Model for Network Intrusion Detection System Using Seq2Seq and ConvLSTM-Subnets," *IEEE Access*, vol. 13, no. January, pp. 30705–30721, 2025, doi: 10.1109/ACCESS.2025.3541399.

[4] V. Sivaprasad, M. Rahmati, S. Member, A. Springer, and H. Vereecken, "Development of Continuous AMSR-E / 2 Soil Moisture Time Series by Hybrid Deep Learning Model (ConvLSTM2D and Conv2D) and Transfer Learning for Reanalyses," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 11169–11182, 2025, doi: 10.1109/JSTARS.2025.3557956.

[5] A. Abdulkareem, H. Brihi, and M. Ghurab, "A Hybrid Deep Learning-Machine Learning Stacking Model for Yemeni Arabic Dialect Sentiment Analysis," *IEEE Access*, vol. 13, no. July, pp. 134160–134171, 2025, doi: 10.1109/ACCESS.2025.3593204.

[6] N. K. Al-qazzaz *et al.*, "Transfer Learning and Hybrid Deep Convolutional Neural Networks Models for Autism Spectrum Disorder Classification From EEG Signals," *IEEE Access*, vol. 12, no. March, pp. 64510–64530, 2024, doi: 10.1109/ACCESS.2024.3396869.

[7] A. Jabbar *et al.*, "A Lesion-Based Diabetic Retinopathy Detection Through Hybrid Deep Learning Model," no. March, 2024, doi: 10.1109/ACCESS.2024.3373467.

[8] J. Ali *et al.*, "A Hybrid Deep Learning Model to Predict High-Risk Students in Virtual Learning Environments," no. July, pp. 103687–103703, 2024.

[9] E. N. Casmiry, R. S. Sinde, and N. M. Mduma, "A Hybrid Deep Learning Model for SQL Injection Attack Detection," *IEEE Access*, vol. 14, no. December 2025, pp. 6450–6463, 2026, doi: 10.1109/ACCESS.2026.3651991.

[10] A. Kala, O. Torkul, and I. H. Selvi, "Early Prediction of Student Performance in Face-to-Face Education Environments : A Hybrid Deep Learning Approach With XAI Techniques," *IEEE Access*, vol. 12, no. December, pp. 191635–191649, 2024, doi: 10.1109/ACCESS.2024.3516816.

[11] C. Hsiao *et al.*, "Precision and Robust Models on Healthcare Institution Federated Learning for Predicting HCC on Portal Venous CT Images," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 8, pp. 4674–4687, 2024, doi: 10.1109/JBHI.2024.3400599.

[12] D. Kumar and A. K. Gupta, "Integrating Data Augmentation and Meta-Heuristic Optimization Algorithms for Enhanced Hybrid Nanofluid Density Prediction Through Machine and Deep Learning Paradigms," *IEEE Access*, vol. 13, no. January, pp. 35750–35779, 2025, doi: 10.1109/ACCESS.2025.3543475.

[13] M. TRIPTI, J. HELEN, S. RASHMI, P. POORVI, A. ABEER, and V. PRAKASH, "Deep vs . Shallow : A Comparative Study of Machine Learning and Deep Learning Approaches for Fake Health News Detection," *IEEE Access*, vol. 11, no. August, 2023.

[14] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, no. October, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[15] S. Roy, "Two-Stage Hybrid Deep Learning Architecture for Cross-Domain Anomaly Detection and," *IEEE Open J. Comput. Soc.*, vol. 7, no. December 2025, pp. 117–128, 2026, doi: 10.1109/OJCS.2025.3643329.

[16] S. MAYANDA MEGA, B. T., J. KASIYAH, and L. OENARDI, "Automatic Detection of Students ' Engagement During Online Learning : A Bagging Ensemble Deep Learning Approach," vol. 12, no. July, pp. 96063–96073, 2024.

[17] C. G. Onyiagha, G. F. Yanwalo, and N. E. Ajimah, "Phishing URL Detection: A Basic Machine Learning Approach," *Int. J. Sci. TECHNOLEDGE*, vol. 12, no. 3, pp. 8–14, 2024, doi: 10.24940/theijst/2024/v12/i3/st2403-001.

[18] M. Hamza *et al.*, "New Hybrid Deep Learning Models to Predict Cost From Healthcare Providers in Smart Hospitals," *IEEE Access*, vol. 11, no. August, pp. 136988–137010, 2023, doi: 10.1109/ACCESS.2023.3336424.

[19] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms Heart disease classification using optimized Machine learning algorithms," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 2, pp. 31–42, 2023, doi: 10.52866/ijcsm.2023.02.02.004.

[20] P. H. Hutagalung, S. Informasi, and U. Nasional, "Heart Disease Classification Using Optimised XGBoost and Random Forest with SHAP Explanations," *Sink. J. dan Penelit. Tek. Inform.*, vol. 10, no. 1, pp. 330–342, 2026, doi: 10.33395/sinkron.v10i1.15544 e-ISSN.

[21] F. Gemci, "A Novel Comparative Approach : Logistic Regression Enhanced by Bat Optimization Versus Logistic Regression Enhanced by Deep Belief Network for Remote Homologous Protein Detection," *IEEE Access*, vol. 13, no. November, pp. 209723–209728, 2025, doi: 10.1109/ACCESS.2025.3641298.

[22] B. Zhang and C. Peng, "Robust Sparse Logistic Regression With the L q (0 < q < 1) Regularization for Feature Selection Using Gene Expression Data," vol. 6, 2018, doi: 10.1109/ACCESS.2018.2880198.

[23] U. Irvine, "Cleveland Heart Disease dataset," Dataset. Accessed: Apr. 04, 2026. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[24] Z. Shangli, Z. Lili, Q. I. U. Kuanmin, L. U. Ying, and C. A. I. Baigen, "Variable Selection in Logistic Regression Model ∗," vol. 24, no. 4, 2015, doi: 10.1049/cje.2015.10.025.

[25] N. Fazakis, S. Kotsiantis, and Y. Dimakopoulos, "MAGA : A Controlled Minority-Class Generative-Augmentation Framework With Fairness-Aware Selection for Imbalanced Diabetes Tabular Classification," *IEEE Access*, vol. 14, no. March 2026, 2026.

[26] Y. Chen, S.-S. LIN, Y. SHI, T.-Y. HO, and X. XU, "MCC : Multi-Cluster Contrastive Semi-Supervised Segmentation Framework for Echocardiogram Videos," *IEEE Access*, vol. 13, no. February, pp. 30543–30554, 2025, doi: 10.1109/ACCESS.2025.3541173.

[27] L. I. U. Chunhui, Q. I. Yue, and D. Wenrui, "The Data-Reusing MCC-Based Algorithm and Its Performance Analysis ∗," vol. 25, no. 4, 2016, doi: 10.1049/cje.2016.06.019.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[28] R. Matsuo, S. Yasuda, and H. Yoshida, "Multiple Instance Learning With Instance-Level Positive-Unlabeled Learning in Anomaly Detection," *IEEE Access*, vol. 13, no. April, pp. 103627–103639, 2025, doi: 10.1109/ACCESS.2025.3578651.

[29] S. K. Patel, S. Member, J. Surve, and G. S. Member, "Encoding and Tuning of THz Metasurface-Based Refractive Index Sensor With Behavior Prediction Using XGBoost Regressor," *IEEE Access*, vol. 10, pp. 24797–24814, 2022, doi: 10.1109/ACCESS.2022.3154386.