Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 4 (2026) 11-20

RESEARCH ARTICLE

Performance Evaluation of Supervised Machine Learning Models for Heart Disease Prediction Using Clinical Survey Data

Avaliação da Performance de Modelos de Aprendizado de Máquina Supervisionado para Predição de Doenças do Coração Usando Dados de Censos Clínicos

Md. Nasim Ali[1]*, Mahir Bin Ayub[1], Sohel Rana[1]

Abstract: This study focuses on evaluating the performance of supervised machine learning models using clinical survey data. We constructed a machine learning pipeline to evaluate the performance of the heart disease prediction models. For this study, we collected clinical data from 399 patients at Cumilla Medical College Hospital. The features included gender, age, blood group, smoking, diastolic pressure, systolic pressure, glucose, hypertension, chest pain (CP), BMI, family medical history (family P), stroke, and heart disease (Heart problem). Heart disease (Heart Problem) was used as the target variable. Initially, we encoded all categorical features, including the binary target label, then separated the target variable. We used Random Over-Sampling to balance the distribution of classes. We applied random over-sampling only on the training dataset. We randomly divided the dataset into 80% training and 20% testing sets. We trained and evaluated five machine learning models on split data, including SVM, Random Forest, Logistic Regression, Decision Trees, and K-Nearest Neighbors. Finally, we calculated accuracy, precision, recall, and the F1-score from the Confusion matrix analysis and 5-fold cross-validation of these models. Then, we compared the evaluation performance of these models. We observed that the Random Forest achieved the highest test accuracy of 90%, and the cross-validation accuracy was 93%. The Support Vector Machine and Logistic Regression achieved slightly lower accuracy. K-Nearest Neighbors had the poorest performance, comparatively lower recall for the positive class. Ensemble models, notably Random Forest, appear to have potential in clinical practice but should be evaluated in larger and more heterogeneous populations. Using clinical data that was acquired locally, this work offers a comparative performance benchmark and highlights generalization behavior across validation methodologies.
Keywords: machine learning — random forest — cross-validation — heart disease — precision — accuracy

Resumo: Este estudo avalia modelos de aprendizado de máquina supervisionado na previsão de doenças cardíacas. Coletamos dados clínicos de 399 pacientes do Hospital Cumilla Medical College, definindo a doença cardíaca (Heart Problem) como variável alvo. As características analisadas incluíram sexo, idade, tipo sanguíneo, tabagismo, pressões diastólica e sistólica, glicose, hipertensão, dor no peito (CP), IMC, histórico médico familiar (family P) e acidente vascular cerebral. Após codificar as variáveis categóricas, dividimos o conjunto em 80% para treinamento e 20% para teste, aplicando o Random Over-Sampling exclusivamente no conjunto de treino para balancear as classes. Avaliamos cinco modelos (SVM, Random Forest, Regressão Logística, Árvores de Decisão e K-Nearest Neighbors) utilizando matriz de confusão e validação cruzada de 5 vias para comparar sua acurácia, precisão, recall e F1-score. O Random Forest obteve o melhor resultado (acurácia de teste de 90% e de validação cruzada de 93%), seguido de perto por SVM e Regressão Logística. O K-Nearest Neighbors apresentou o pior desempenho, com um recall notavelmente menor para a classe positiva. Assim, modelos ensemble, especialmente o Random Forest, mostram grande potencial clínico, embora exijam testes em populações maiores e mais heterogêneas. Utilizando dados adquiridos localmente, este trabalho fornece um benchmark comparativo e destaca o comportamento de generalização em diferentes metodologias de validação.
Palavras-Chave: aprendizado de máquina — random forest — validação cruzada — doença cardíaca — precisão — acurácia

[1] *Electrical and Electronic Engineering, CNN University of Science and Technology (CCNUST), Bangladesh*
***Corresponding author**: ali.nasim.212212@gmail.com
DOI: http://dx.doi.org/10.22456/2175-2745.153722 • **Received:** 19/02/2026 • **Accepted:** 02/06/2026

1. Introduction

Machine learning has a significant impact on predictive analysis for real-world data. It enables algorithms to locate patterns and generate predictions. Machine learning models might also be supervised, unsupervised, or reinforcement learning-based. Meanwhile, supervised models learn from labeled data, and unsupervised models discover patterns in unlabeled data. Reinforcement learning enables agents to learn how to act based on rewards. ML is used in various domains, including healthcare, finances, and autonomous driving[1]. Heart and cardiovascular disease are the first causes of death worldwide; for example, aging, high blood pressure, diabetes, smoking, obesity, or high cholesterol increase the risk. The issue with CVD risk is that it accumulates silently, and many people don't notice until they see serious events such as heart attacks or strokes. Thus, early identification of these at-risk patients is essential to avoid such complications and enhance their outcome [2]. Conventional CVD risk assessment relies on history, physical examinations, laboratory tests, and imaging. We follow the methods of statistical models/clinical scoring for screening in the population. These approaches may be inaccurate due to data issues such as missing values and population bias [3]. Today's healthcare systems produce large volumes of structured data from various sources, such as EHRs (demographic, vital signs, laboratory results, and medications). This allows machine learning (ML) models to detect intricate patterns within patient data that conventional models may overlook. ML can model nonlinear relationships and interactions between risk factors that simple models cannot express. Numerous studies have reported that ML models achieve high predictive performance in CVD, but the results are inconsistent according to the characteristics of the data set and the validation method [4]. Although the results are encouraging, ML models often have limitations such as dataset shifts and no external validation. All the hype will fade when we develop ML models that outperform actual clinical decision-making on outcomes relevant to patients. Conventional models, such as logistic regression (LR), k-nearest neighbors (kNN), and decision tree (DT), are still commonly used for CVD prediction because they are simple and interpretable. However, there is a general belief that ensemble-based approaches, such as Random Forest (RF), produce superior generalization and performance [5][6]. Feature selection is also of paramount importance. In clinical data, features may have high correlation or be weak predictors, causing overfitting and lack of interoperability. Choosing the most intersectional features is necessary to enhance performance and interoperability [7][8]. It is difficult to outperform classical models such as RF, SVM, and LR without sophisticated tuning or additional data. Nevertheless, these models serve as a satisfactory benchmark to set the standard in clinical predictions. This concept is particularly relevant in health care. Patients and providers will need to believe the forecasts, and the models will have to be understandable. By integrating ML with explainability techniques, we may be able to improve the performance and interpretability of CVD prediction models[9]. In this paper, we compare traditional ML algorithms (kNN, LR, SVM, DT, and RF) with the same dataset and evaluation measures for performance evaluations in predicting heart disease. We have not intended to identify heart disease. Rather, we sought to estimate how these models will perform in a clinical setting. This benchmark is a significant reference point because the literature indicates substantial variability in performance arising from differences in processing, feature processing, and validation procedures[10][11].

2. Related work

The Random Forest (RF) algorithm is one of the most successful machine learning services, as it is highly flexible and capable of handling almost all types of input data. But it can be bad for small datasets. The studies considered in this paper focus on the use of Random Forest for small datasets with MPDL, specifically medical and text classification. They demonstrate the strengths and weaknesses of Random Forest and propose heuristics to make it more successful on small data. Han, Williamson and Fong (2021) have examined the power of RF on limited clinical datasets, including vaccine trial data. They observed that Random Forest does not perform well with small data sizes. They may also introduce overfitting. Simpler models like Logistic Regression and Naive Bayes perform better than Random Forest, which was only 94% correct. The report also recommended that Random Forest is talked about a lot, but it might not work well to small datasets, as it overfitting. With small datasets, overfitting is a danger. They felt that applying different methods for selecting features would lead to improving the accuracy of the Random Forest. Their article does not contain hybrid representations such as Random Forest plus other strategies. Cross-validation methods were also not employed. We could have used it to make an improved guess about the new information. Later researchers may study the combination of the Random Forest method with other algorithms to enhance its accuracy and apply it to small data samples [12] Govindu and Palwe (2023) They collect the MDVP vowel-phonation voice dataset with 31 subjects, of whom 23 have PD, and experiment on 195 recordings. For each recording, there are 22 acoustic features and a class label (0 = healthy, 1 = PD). They also have unbalanced partitioned data, consisting of 109 recordings from subjects with PD and 40 recordings from normal subjects. They up sample the normal class to be 109 vs 109, which does not equal "195 recordings" (the abstract says "30" people). They contrast four classifiers: LR, RF, SVM and KNN. They use StandardScaler, a 75/25 train–test split, and three settings (with all 22 features, PCA to 5 features, and balanced (up sampled) data). The highest accuracy achieved is approximately 91.83%. Random Forest is the best with all the features (91.83%), SVM is the best post-PCA (91.75%), and KNN matches the best with balanced data (91.83%). In total ,the article reports that MDVP's voice-based features can accurately support diagnosis of Parkinson's at around 92%

accuracy. This indicates that voice recordings may be useful for Telemedicine style early screening, and Random Forest (and in some cases SVM/KNN) is most effective [13].

Uddin,N et al. (2026) employed a Random Forest model to predict heart disease based on the UCI Heart Disease dataset. Each sample has 303 measurements and 14 "features"—such as age, cholesterol level, or blood pressure. They intended to use these clinical measures to predict if a patient had heart disease. The authors compared Random Forest with other models and SVM as well as Logistic Regression. The model that yielded the highest accuracy using hyperparameter tuning was Random Forest with 93.41%. This study showed that Random Forest can deal with small clinical datasets properly provided that fine-tuning is performed. The Small sample size is a common problem in sample size in machine learning. This may cause overfitting. The problem of class imbalance was also not solved in this study. Medical imaging datasets are typical in this respect. Classes must be balanced by SMOTE like strategies and other studies that enhance model performance [14]. In a different work, Parami, N. et al. (2024) predict heart disease. They applied it to the Cleveland Heart Disease dataset, which consists of 303 samples. They also compared Random Forest with SVM, K-NN, and Logistic Regression. Random Forest can achieve up to 94% accuracy. The study revealed that Random Forest is the best for small datasets and has perfect performance on clinical data. But it brought back the class imbalance issue. Parami,N et al. investigated the imbalance of this representation and the subsequent effect on Random Forest, among other methods. The Random Forest classifier worked well for us, but it would be interesting to try using oversampling methods like SMOTE or ADASYN as a solution for the imbalance in data. The generalizability of the model to other clinical databases is untested. This paves the way for further investigations into its performance [15]

Rani,P. et al(2021). introduce a machine-learning-based decision support system (DSS)[16]. They determine whether a patient has or does not have heart disease from clinical data. The UCI (Cleveland Heart Disease) dataset, which was applied in the work. Has a total of 14 features (8 categorical and 6 numeric variables). Before applying SMOTE to alleviate imbalance, the class distribution is 164 samples in class 0 (no heart disease) and 139 samples in class 1 (heart disease). Their pipeline imputes missing values with MICE. With the hybrid method of a genetic algorithm and recursive feature elimination (GA–RFE), it performs selection on features. They solve class imbalance by adding the SMOTE. They scale input data using standard scaling before training any models,s well. They learn and compare multiple classifiers, including logistic regression, SVM, Naive Bayes, Random Forest, and AdaBoost. Random Forest has the highest accuracy rate of 86.6%. The DS authors describe the DSS as a tool forlornly screening and clinical decision-making. It does, however, only forecast disease presence and not strength. They make some recommendations for future work that include more aggressive optimization techniques and deep learning. They also suggest trialing the system in actual clinical practice

Table 1. : Related work using short datasets

Authors	Published Date	Dataset	Models Used	Best Accuracy (RF)
S. Han, B. D. Williamson, Y. Fong	2021	Small Clinical Datasets (e.g., vaccine trials)	Random Forest, Logistic Regression, Naive Bayes	93.99%
Rani,P et al	2021	dataset (UCI), 164 (class 0) and 139 (class 1) before SMOTE balancing.	Random Forest, Logistic Regression, SVM	92%
Govindu and Palwe	2023	voice dataset with 31 subjects, whom 23 have PD, 195 recordings each have 22 features.	Random Forest, LR, SVM and KNN.	91.83%
N. Parami, R. Khapre, P. Mehra, V. Rao, A. Chhabria	2024	Cleveland Heart Disease Dataset (303 samples)	Random Forest, SVM, KNN, Logistic Regression	94%
This study		clinical survey data of 399 patients from local hospital	Random Forest, Logistic Regression, SVM, DT, kNN	test accuracy 90% , and Cross-validation accuracy 93%

3. Methodology

As shown in Figure 1, the main methodology of this study follows a machine learning pipeline schematic to evaluate the performance of Machine Learning models and ensure comparison. This pipeline has eight sequential stages, such as Input data, Data Preprocessing, Target Label separation, Train-Test split, Data Balancing, Training Data Processing, Model Training, and Evaluation.

3.1 Data source

Data for the study were derived from primary data collected in a local government hospital. The hospital was Cumilla Medical College Hospital, Bangladesh. A structured questionnaire was used to collect data from patients hospitalized in the

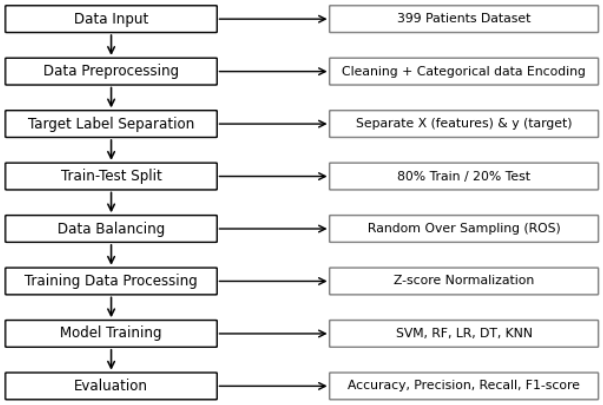

Figure 1. Block diagram of proposed machine learning pipeline

cardiology wards during the study period (13 July 2025 to 01 November 2025) and was supplemented by clinical measurements at admission. Only adult patients (18 years or older) for whom all relevant variables of interest were complete and with prior-known diagnoses indicating the presence or absence of heart disease were retained, while records with missing outcome tags or major missing predictor values were excluded. The target label (Heart Problem) was identified using patient questionnaire responses and hospital clinical records obtained during the survey procedure at Cumilla Medical College Hospital. The last data set consisted of 399 patients (yes with heart disease and no without), in which 13 attributes were documented for each patient: sex, age, blood group, smoking, diastolic pressure, systolic pressure, glucose, hypertension, chest pain (CP), BMI, Family Medical History (Family P), stroke, and Heart Disease (Heart Problem). All data were anonymized prior to analysis and the study was conducted according to current ethical standards, as approved by the responsible institutional committee.

3.2 Data Preprocessing

Processing was performed in several steps before the model was trained. First, we conducted data cleaning to verify manually aggregated survey records, remove duplicate respondents, and address obvious errors or missing values in predictor variables. The rest of the variables were then analyzed in order to find numerical characteristics (i.e., age, diastolic and systolic blood pressure, glucose, BMI) and categorical features (i.e., gender, blood group, smoke status, hypertension, chest Pain (CP), Family Medical History(family p), and stroke). The binary categorical outcome for the goal variable, Heart problem, indicated whether or not heart disease was present. Categorical variables were transformed to numeric data by applying the proper label encoding. We then separated the target label (the presence or absence of heart disease) from the feature matrix to make a clear distinction between input features and outcome. Here numerical input features were scaled to a unit scale so as to enable better performance for distance-based and margin-based learning approaches like k-nearest neighbors and SVM.

3.3 Training-Test Data split

To evaluate the performance and the generalization ability of the trained ML models, we classify the hospital data into a training partition and a testing partition. This is crucial so that developed models should be able to predict heart disease not only for known cases but also for unknown patients[17]. The dataset was divided into 80:20, with 80% of the data for training and the rest for testing. To spilt data, we used random state-42.The training dataset provides patterns for the model to learn and one can find all kinds of patterns/relations between input features and target class. Subsequently, the internal parameters within the model are adapted to minimize the prediction mistake. During training, the test samples are not used. It is only used to test the model after its training. This technique could be useful in determining which models were more robust and predictive and would reduce the possibility of overfitting. The generalization of a model to real clinical situations may be tested by testing it on unknown data. All the models, DT, RF, KNN, LR, and SVM of machine learning are trained and tested over the same data split for fair comparison. The performance of each model is reported under the standard evaluation metrics: accuracy, precision, recall and F1-score[18][19].

3.4 Training data Oversampling

We adopted a data-balancing technique (Random oversampling) on the training set to compensate for class imbalance and prevent bias towards the majority class. As a result, no synthetic samples were generated, and the test set remained imbalanced, which meant the test set was realistic.

3.5 Processing training Data

Input numerical features were scaled by using the StandardScaler technique to enable better performance for distance-based and margin-based learning approaches like k-nearest neighbors and SVM. There were twelve input features and one target feature in our dataset. All 12 input features were used for model training. The same input features were utilized as input for five classifiers (logistic regression, SVM, k nearest neighbors, random forest, and DT) to keep the comparison fair and uniform among model performances.

3.6 Correlation Matrix Analysis

Following Figure 2, all dummy-encoded categorical variables and the numerical features exhibit the above correlation. The moderate size correlation (0.75) between SMOKING and GENDER indicates that there is an important association between gender and smoking status, so that a particular sex may likely have different rates of smoking than the other. Diastolic Pressure and Systolic Pressure are highly related (0.69), demonstrating that these two blood pressure readings are somewhat dependent (they will generally move in the same direction for any one individual). H8 (HYPERTENSION) is

moderately related to both Diastolic Pressure (0.69, Figure 2) and Systolic Pressure (0.68, Figure 2), following our intuition about the impact of high blood pressure on these two measurements. The constant Chest Pain (CP) has a high positive correlation of 0.67 to Heart Disease (HEART PROBLEM), indicating the significant role taken by chest pain in predicting heart problems. Moreover, FAMILY P (family history of heart disease) and STROKE have moderate associations with HEART PROBLEM (0.22 and 0.35, respectively), indicating a heritability as well as possibly an underlying cardiovascular link because family history is a risk factor for heart disease. Overall, it is the correlation matrix that yields some clues about which clinical indicators likely have an effect on heart disease and other health outcomes. These results may support feature importance analysis in machine learning models used to predict cardiovascular diseases.

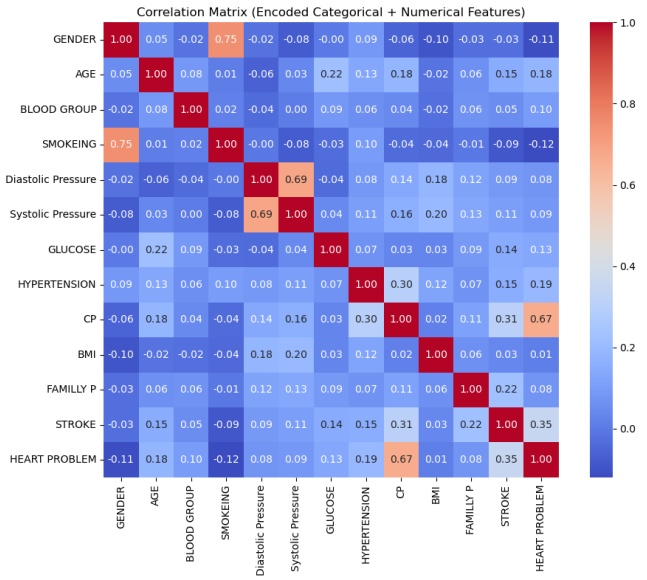

Figure 2. Correlation matrix

3.7 Cross-validation

Cross-validation is a cornerstone of model performance evaluation and model selection in machine learning, taking place by providing an accurate estimate of how well the predictive model can be generalized to new or unseen data. Instead of having one train/test split, cross-validation performs a pair of disjoint partitions on the dataset in a systematic way (folds) and iteratively trains the model on some subset of the data before testing it on another subset, so that eventually all data points get sampled for both training and testing at least once. This procedure reduces overfitting, controls the bias and variance in the assessment results, and enables valid comparison of models/configurations by averaging performance values obtained for different folds. Popular variants involve k-fold cross-validation, where a dataset is split into k equal parts, and stratified cross-validation, which maintains the class distribution in classification tasks. More sophisticated methods, such as nested cross-validation or time-series cross-validation, further develop the plain framework so as to accommodate hyperparameter tuning and temporal dependencies. The particular type of cross-validation used in a study will depend on the nature of the dataset and the modeling goals of that study, but in every case, its application increases confidence that the reported performance reflects a model's actual predictive ability rather than accidental properties of one data split[20]. We used 5-fold cross-validation on training data. Scaling and oversampling were applied on training data in each fold to prevent information leakage. The model and preprocessing hyperparameters were the same for every fold. There is no independent tweaking done for each fold only the training/validation split varies.

3.8 Training and Testing Models

The machine learning models offer numerous data analysis processes to medical diagnoses and clinical decision-making, which are correlated to both advantages and disadvantages. The five best models have been defined, as they can be applied in the medical fields, including the determination of the disease, the forecasting of the prognosis, and the combination of the multimodal data. They are support vector machines (SVMs), decision trees (DTs), logistic regression (LR), k-nearest neighbor (KNN), and random forest (RF)[21].

3.9 Random Forest classifier

Random Forest (RF) is another ensemble learning method, and in this case, it utilizes very many decision trees, which are learned on bootstrap samples, and random feature sets utilized in the process to minimize overfitting and enable resilience. RF can be applied to any medical data, including imaging characteristics and clinical variables, and the ranks of feature importance can be used to support the interpretation of the results. Decisions made by individual trees, on the other hand, are not as transparent. RF is an overall baseline model, which proves to be widely applied in patient diagnosis and prognosis and is often combined with feature selection or calibration procedures to enhance its accuracy between subgroups of patients. The representative's work also includes the application of RF in medical classification, the processing of real-life data, and a focus on calibration[22].

3.10 Support Vector Machine (SVM)

SVMs perform well in large dimensional space when the sample size of features is lower than that of the samples. This property is because SVMs maximize the separation between classes, and this maximization is achieved through the use of a nonlinear method with the assistance of a kernel, which contributes to SVMs' effectiveness in terms of generalization. Such properties have made SVMs relevant to radiometric and genomics. However, their practical usefulness is limited in some cases due to challenges such as selecting the appropriate kernel, tuning the kernel parameters, processing large datasets, and explaining high-dimensional decision boundaries to clinicians. The other larger studies demonstrate the

operations of SVMs medication as well as their application in the multi-omics connection[23][24].

3.11 Logistic Regression (Logistic-Reg)

The reason Logistic Regression (LR) is considered an outdated model is that it is easy to understand, clear, and easy to interpret. Logistic regression is particularly suitable for categorical classification, especially when there is an approximately linear dependence between the characteristics and the results. The rationale for this is that the clinical decision-making regarding the LR coefficients as a risk factor is supported by its interoperability, the fixed calibration characteristics of the model, and its potential as a baseline. It is not so effective, however, when the interactions among features are more complex and nonlinear. LR has been shown to be important in some of the most impactful studies as a risk stratifier and a regulatory-friendly environment over the more complicated models[25][26]. It is a statistic model to processes the binary classification problems. It makes a prediction of the probability that a given input is most likely to be in one class. Its name is a misnomer: logistic regression isn't for regression, it's for classification. Based on the logistic or sigmoid function which transforms any input into a value between 0 and 1, you can't get your probability outside those limits.

3.12 k-Nearest Neighbors (KNei / KNN)

One such technique is the instance-based technique k-Nearest Neighbors (KNN), which is simple to apply and might be convenient in the low dimension of the classes when the classes can be easily told apart. KNN is susceptible to large-scale computational costs, feature sensitivity on feature scale, high-dimensionality sensitivity, and high-dimensional clinical data. The curse of dimensionality is detrimental to KNN at prediction time. Despite these inadequacies, KNN is an inspirational source model of medicine diagnosis as well as learning. It too possesses some significant writings regarding the medical performance and its issues with high-dimensional space[27][28].

3.13 Decision Trees (DTs)

The decision trees (DTs) provide a rule-based decision-making approach and can be depicted and interpreted graphically in a very simplistic way, which is why they became one of the best known decision-making methodologies in the education outline and in clinical practice. It surpasses single decision trees, which overfit and struggle with noisy data, and it yields interpretable results. Decision trees are commonly used as a foundation for straightforward procedures in diagnostics; however, ensemble techniques are generally more predictive. Representative works[29][30].

3.14 ML Models Hyperparameter

A uniform set of hyperparameter was applied to every experiment in order to provide an equitable and repeatable comparison between various machine learning models. With the exception of a few instances where modification was required to enhance performance, the majority of the hyperparameters were left at their default values when the models were created using the Scikit-learn module.Table 2 provides a summary of all the hyperparameter settings used in this investigation.

Table 2. Hyperparameter settings of machine learning models

Model	Hyperparameters
SVM	kernel=RBF, C=1.0, gamma=scale, probability=True, random_state=42
Logistic Regression	max_iter=1000, penalty=L2 (default), solver=lbfgs (default), random_state=42
Decision Tree	criterion=gini, max_depth=None, min_samples_split=2, random_state=42
KNN	n_neighbors=5, metric=minkowski, weights=uniform (default)
Random Forest	n_estimators=100, criterion=gini, bootstrap=True, random_state=42

3.15 Evaluation metrics

The confusion matrix in Figure 3 is based on four elements, such as True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Values are tracked for control of some The accuracy is the percentage of correctly

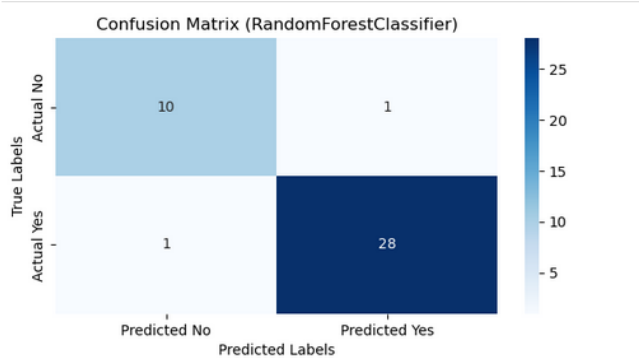

Figure 3. Confusion matrix

classified samples to total number of samples It is calculated as:

Precision is a measure to ensure the trustworthiness of positive predictions, and is presented as the ratio of true-positive cases with respect to those positively estimated[31]. The accuracy is the percentage of correctly classified samples to total number of samples It is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Precision is a measure to ensure the trustworthiness of positive predictions, and is presented as the ratio of true-positive cases with respect to those positively estimated[31].

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Recall, also known as sensitivity is employed to evaluate the performance of the classifier in detecting the true positive cases [32]

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

The F1-score is the harmonic mean of recall and precision, and provides a more balanced consideration regarding the performance of the classifier.

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

4. Results and Discussion

The classification results of five machine learning models with confusion matrices are shown in Figure 4. Each confusion matrix accounts for true negatives (TNs), false positives (FPs), false negatives (FNs), and true positives (TPs). Instead of using the entire dataset for evaluation, the held-out test set was used. There is a moderately imbalanced test distribution in favor of the positive class. For this reason, the confusion matrices are invaluable here, as they show what is correct and incorrect not just in terms of total accuracy but also what kind of errors each model makes (FP vs FN), which is critical when applied to many real-world decision systems. Random oversampling was applied to the training dataset after the data split to reduce information leakage. Here, the test dataset was retained in its original unbalanced form. As a result, the reported confusion matrices reflect the original model's generalization. The SVM model obtained TN = 17, FP = 4, FN = 6, and TP = 53. This leads to an accuracy of 87%, high precision of 87%, a recall of 87%, and an F1-score of 87%. The error profile reveals that the SVM model produced 4 false positive predictions, indicating that there are not many negative samples that are classified as positive. However, this test resulted in 6 false negatives, indicating a small proportion of true positive cases were undetected. These results indicate that SVM achieved balanced classification performance. Logistic Regression achieved TN = 18, FP = 3, FN = 7, and TP = 52, resulting in an accuracy of 87%. Logistic Regression produced 3 false positives, which was lower than Decision Tree and KNN. It led to a precision of 88% among the tested models. However, it generated 7 false negatives; thus, the recall dropped slightly (87%). The F1-score (87%) is still high, which suggests that Logistic Regression will make conservative positive predictions (very few FP) but miss a small number of positives. The TN, FP, FN, and TP for the Decision Tree were 14, 7, 6, and 53, respectively, with an accuracy of 83%. The Decision Tree produced 6 false negatives, which was comparable to SVM and slightly lower than Logistic Regression. These findings led to a decrease in recall (83%) and F1-score (83%). The confusion matrix shows that the Decision Tree demonstrated moderate classification performance but generated comparatively higher false positive predictions; it may miss some positive instances, possibly due to an overfit of decision boundaries from a single tree. KNN obtained TN = 15, FP = 6, FN = 8, and TP = 51, with the worst classification result (accuracy 82%). The recall is 82%, resulting in an F1-score of 82%. The precision remained 83%. This result implies that KNN might be sensitive to local neighborhood structure and feature scaling or need more advanced tuning of k and distance measures, especially in class-imbalanced conditions. Of all the models, the best results were achieved by Random Forest. It reported the best accuracy (90%), recall (90%), and F1-score (90%). Crucially, Random Forest had the minimum number of false negatives (4) and therefore possibly the best ability to accurately classify positive/target cases. While its false positives (4) were similar to SVM but lower than Decision Tree and KNN, it reduced false negatives more effectively in comparison with SVM and Decision Trees, implying that the ensemble structure increases sensitivity without over-penalizing FP. In general, we can see from the confusion matrix that Random Forest offers a better error trade-off and is especially good at reducing false negatives if missing positive cases is not desirable. Overall, the evaluated models demonstrated relatively high precision values, indicating reliable positive predictions, i.e., when a model predicts a positive class, it is mostly true. The major disparities are observed in false negatives and specificity that have a direct impact on recall as well as negative predictive value. The Random Forest algorithm had the least number of total errors (8 misclassifications) and also among the fewest false negatives (4 false negatives), making it the best overall performer. SVM and Logistic Regression made up a second tier with 9 total errors each, in which the performance of both methods was similar in recall, but SVM and Logistic Regression achieved similar overall performance, although Logistic Regression produced slightly fewer false positives. Decision Tree and KNN exhibited lower overall performance due to higher misclassification rates. Because there are more positive instances than negative instances in the test set (class imbalance), accuracy should not be the only consideration. Hence, the patterns revealed by the confusion matrix and their associated measures (recall and F1-score) indicate this. In general, Random Forest is the best model, although in cases where identifying positive samples accurately is more important while having an interest in minimizing false alarms (FP), other models such as Logistic Regression might be preferable.

The following Table 3 compares the five machine learning models to each other in two evaluation modes: train–test split accuracy and mean accuracy from cross-validation. The cross-validation mean is typically higher than the single test. This implies that cross-validation provides a more stable and less variance-sensitive estimate. The SVM model achieved a test accuracy of 0.87 and a cross-validation average of 0.87. This suggests that SVM works well in both the split data and 5-fold cross-validation. All—precision, recall, and F1-score—are around 0.87, showing that they maintain a balance between

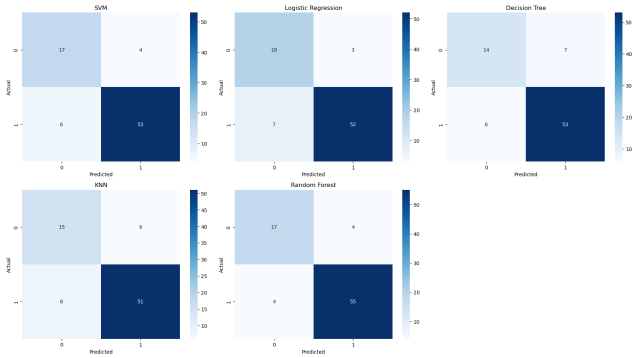

Figure 4. Performance of five models

correct positive predictions and missed cases. Logistic regression gets a 0.87 test accuracy and a 0.84 cross-validation mean as well. It has a slightly higher precision (0.88) than SVM, resulting in an F1-score of 0.89, and recall of 0.87. This implies that Logistic Regression has a higher probability of predicting the positive class correctly as compared to the negative class (i.e., it causes fewer false positives and more true positives) but not at the cost of increasing the false negatives. The decision tree has 0.83 test accuracy and a 0.90 cross-validation mean. It has a precision of 0.83, recall of 0.82, and F1-score of 0.83.Although its held-out test accuracy was lower, the Decision Tree achieved a higher cross-validation mean accuracy (0.90) than SVM (0.87) and Logistic Regression (0.84). The KNN model showed comparatively lower performance, achieving a test accuracy of 0.82 and a mean cross-validation score of 0.80. Its precision (0.83), recall (0.82), and F1-score (0.82) were also the smallest of all the models. This result shows that KNN does not do a good job of separating classes for this dataset, and its performance is not always consistent across folds. Random Forest achieves the highest value for each method in Table 5. It has the maximum test accuracy (0.90) and its cross-validation mean is also the highest (0.93). All accuracy, recall, and F1-score values were 0.90, demonstrating steady and dependable classification performance throughout folds.That being said, Random Forest has the best trade-off between sensitivity (detecting the positive class) and specificity (avoiding wrong predictions). In general, RF is the most accurate and the most stable model of these five methods. SVM and Logistic Regression are the second-best group in terms of accuracy, with comparable test performances. As shown in Figure 5, the model depicts accuracy for five different models across five folds (5-fold cross-validation (CV)). Each color and marker indicates the different performance of a model. The SVM (blue) exhibits relatively stable accuracy, with gentle bumps after each fold update, especially in folds 1–4. Logistic Regression (green) shows small fluctuations among folds, with gentle bumps after each update. Decision Tree (red) shows higher fluctuation compared to other models, especially in the fourth fold. For KNN (cyan), there is more fluctuation, with a small plunge in accuracy at some of the folds. The Random Forest (purple) also achieved the smallest

Table 3. ML models performance comparison using Train-Test split vs Cross validation

Model	Test Accuracy	5 Fold Cross-validation Accuracy (CV mean)	Precision	Recall	F1-Score
SVM	0.87	0.87	0.87	0.87	0.87
Logistic Regression	0.87	0.84	0.88	0.87	0.87
Decision Tree	0.83	0.90	0.83	0.83	0.83
KNN	0.82	0.80	0.83	0.82	0.82
Random Forest	0.90	0.93	0.90	0.90	0.90

fluctuation between the folds, so its overall performance was the best. We can see that Random Forest accuracy is higher than other models in each fold. The best performance was

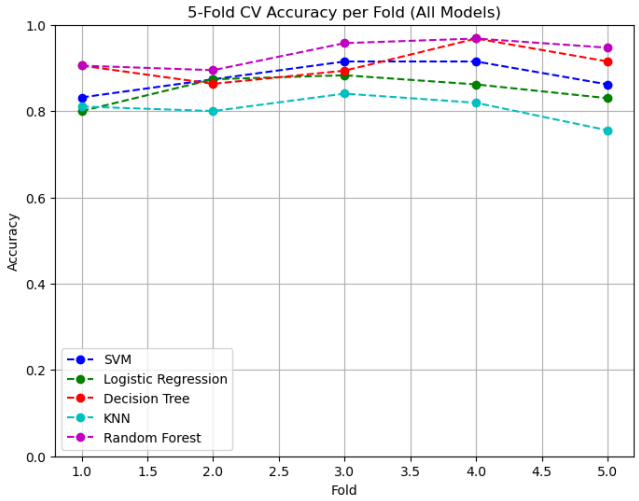

Figure 5. 5-fold cross validation accuracy per fold of all model

observed using Random Forest, with test accuracy 0.90 and cross-validation accuracy 0.93. It achieved high precision, recall and F1-score of 0.90 for positive as well as negative classes. It is very effective because it involves multiple Decision Trees and thereby reduces the overfitting of Data and will increase the generality. It is a preferable strategy commonly employed in complicated tasks such as medical prediction. As a result of performance averaging across several folds, cross-validation accuracy is generally higher than that of a single train-test assessment. Instead of showing that there are no noisy or borderline samples, this averaging produces a more stable estimate of model generalization by reducing variance brought on by a particular data split.

5. Conclusion

In this study, a locally collected dataset from Cumilla Medical College Hospital, Bangladesh, was considered to assess five supervised learning classifiers, such as k-Nearest Neighbour (k-NN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). Both of these models were trained on the same data and tested using various metric types that tended to favor a positive class (heart disease), including accuracy, precision, recall, and the F1-score. However, we observed that the Random Forest model achieved the highest accuracy among other models. Overall accuracy was 0.93 for 5-fold cross-validation with high precision and recall for the positive class and a pretty excellent job in the negative class. This conclusion is in line with other work that found ensemble models (such as Random Forest) achieved high performance when predicting CVR, especially with tabular clinical data and complex feature interactions. In comparison, the k-NN model performed worse. The k-NN model had poor recall for positive cases, indicating it struggled with high-dimensional clinical features, especially when combining text and numeric data. There are, nevertheless, certain restrictions on the present preliminary study. The model was trained on one hospital with a minimum of 399 patients. The small sample size led to unclear generalization performance of the model when applied to other populations. To maximize generalization in future studies, this model will need to be validated on a larger and more diverse sample comprised of patients collected from multiple hospitals or healthcare systems. In addition, although we conducted data pre-processing and training data processing to avoid overfitting and improve the model performance, critical content-specific information is not contained in the dataset. For example, genetic data or the history of the patient in the long term was not used. This model could be enhanced by incorporating such information. Future research could explore whether including such covariates would improve model fit. A second challenge with Random Forests is that, even though it performed very well, the model is not interpretable. Healthcare workers at treatment centers should foster trust by being transparent about their decision-making process. One way to mitigate this disadvantage could be by employing techniques from explainable AI (XAI), such as SHAP or LIME. These strategies could potentially lead to a more interpretable model and clearer decision reasoning for clinicians, which could aid clinical decision support. Finally, this study advances the current body of knowledge regarding the prediction of cardiac disease using machine learning. It is additional to a few other studies that compare various machine learning models regarding cardiovascular risk prediction. The paper critiques avenues for reform, such as validation of the model and selection and interoperability of the features. Further research is needed to implement these models in clinical practice, gather high-quality data that can improve prediction accuracy, and create interpretable solutions that ensure these predictions are actionable in both healthcare and public health sectors.

Acknowledgements

We are deeply grateful to the Comilla Medical College, Cumilla, Bangladesh administration for allowing us to use their clinical survey data for this research. The study was carried out in compliance with hospital policies, fully respecting patient privacy and ethical norms.

Author Contributions

Contributions to this paper were equally made by Md. Nasim Ali, Mahir Bin Ayub, and Sohel Rana. Every author contributed to the study's conception and design, data collecting and preprocessing, machine learning model creation and evaluation, results analysis and interpretation, and manuscript writing and revision. The final draft was read and approved by all authors.

References

[1] SUBASI, O. et al. The landscape of modern machine learning: A review of machine, distributed and federated learning. *arXiv preprint arXiv:2312.03120*, 2023.

[2] A, R. G. et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the american college of cardiology*, Elsevier, v. 76, n. 25, p. 2980–2981, 2020.

[3] WARD, A. et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ digital medicine*, Nature Publishing Group UK London, v. 3, n. 1, p. 125, 2020.

[4] KRITTANAWONG, C. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 16057, 2020.

[5] ALI, M. M. et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, Elsevier, v. 136, p. 104672, 2021.

[6] SUBRAMANI, S. et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in medicine*, Frontiers Media SA, v. 10, p. 1150933, 2023.

[7] PATHAN, M. S. et al. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, Elsevier, v. 2, p. 100060, 2022.

[8] ALFADLI, K. M.; ALMAGRABI, A. O. Feature-limited prediction on the uci heart disease dataset. *Computers, Materials & Continua*, v. 74, n. 3, 2023.

[9] MIENYE, I. D.; JERE, N. Optimized ensemble learning approach with explainable ai for improved heart disease prediction. *Information*, MDPI, v. 15, n. 7, p. 394, 2024.

[10] BANERJEE, A. et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC medicine*, Springer, v. 19, n. 1, p. 85, 2021.

[11] LIU, T. et al. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal-Digital Health*, Oxford University Press UK, v. 6, n. 1, p. 7–22, 2025.

[12] HAN, S.; WILLIAMSON, B. D.; FONG, Y. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC medical informatics and decision making*, Springer, v. 21, n. 1, p. 322, 2021.

[13] GOVINDU, A.; PALWE, S. Early detection of parkinson's disease using machine learning. *Procedia Computer Science*, Elsevier, v. 218, p. 249–261, 2023.

[14] UDDIN, N. et al. Hpml-cvd: hyperparameter tuned machine learning model to predict cardiovascular disease. *Biomedical Materials & Devices*, Springer, v. 4, n. 1, p. 861–875, 2026.

[15] PARAMI, N. et al. Predicting heart disease using random forest: A comparative analysis. In: SPRINGER. *International Conference on Computer Vision and Robotics*. [S.l.], 2024. p. 15–25.

[16] RANI, P. et al. A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, Springer, v. 7, n. 3, p. 263–275, 2021.

[17] LAMIR, A. A.; RAZZAGZADEH, S.; REZAEI, Z. A comprehensive machine learning framework for heart disease prediction: Performance evaluation and future perspectives. *arXiv preprint arXiv:2505.09969*, 2025.

[18] CHANDRASEKHAR, N.; PEDDAKRISHNA, S. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, MDPI, v. 11, n. 4, p. 1210, 2023.

[19] MOHAMADI, Z. et al. The application of random forest-based models in prognostication of gastrointestinal tract malignancies: a systematic review. *Frontiers in Artificial Intelligence*, Frontiers Media SA, v. 8, p. 1517670, 2025.

[20] ALLGAIER, J.; PRYSS, R. Cross-validation visualized: a narrative guide to advanced methods. *Machine Learning and Knowledge Extraction*, MDPI, v. 6, n. 2, p. 1378–1388, 2024.

[21] HONG, W. et al. Usefulness of random forest algorithm in predicting severe acute pancreatitis. *Frontiers in cellular and infection microbiology*, Frontiers Media SA, v. 12, p. 893294, 2022.

[22] OMANKWU, O. C.; OSODOEKE, E. C.; UBAH, V. I. Disease prediction using random forest machine learning algorithm. *NIPES-Journal of Science and Technology Research*, v. 6, n. 1, 2024.

[23] LI, H.-J. et al. Radiomics-based support vector machine distinguishes molecular events driving the progression of lung adenocarcinoma. *Journal of Thoracic Oncology*, Elsevier, v. 20, n. 1, p. 52–64, 2025.

[24] GUIDO, R. et al. An overview on the advancements of support vector machine models in healthcare applications: a review. *Information*, MDPI, v. 15, n. 4, p. 235, 2024.

[25] HU, Y. et al. Beyond comparing machine learning and logistic regression in clinical prediction modelling: Shifting from model debate to data quality. *Journal of Medical Internet Research*, JMIR Publications Toronto, Canada, v. 27, p. e77721, 2025.

[26] MANN, J. et al. Machine learning or traditional statistical methods for predictive modelling in perioperative medicine: A narrative review. *Journal of Clinical Anesthesia*, Elsevier, v. 102, p. 111782, 2025.

[27] UDDIN, S. et al. Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. *Scientific Reports*, Nature Publishing Group UK London, v. 12, n. 1, p. 6256, 2022.

[28] HALDER, R. K. et al. Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, Springer, v. 11, n. 1, p. 113, 2024.

[29] COSTA, V. G.; PEDREIRA, C. E. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, Springer, v. 56, n. 5, p. 4765–4800, 2023.

[30] YILDIZ, A. Y.; KALAYCI, A. Gradient boosting decision trees on medical diagnosis over tabular data. In: IEEE. *2025 IEEE International Conference on AI and Data Analytics (ICAD)*. [S.l.], 2025. p. 1–8.

[31] BETRU, B.; GETAHUN, F. Ontology-driven intelligent it incident management model. *IJITCS*, v. 15, p. 30–41, 2023.

[32] SUHARJO, B.; INDRAJIT, R. E. et al. Performance comparison of k-nearest neighbor, decision tree, and random forest methods for classification of cyber defense master scholarship recipients. *Internet of Things and Artificial Intelligence Journal*, v. 4, n. 4, p. 666–676, 2024.