www.nature.com/scientificreports

scientific reports

OPEN

Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy

Yagyanath Rimal[1,2]✉, Navneet Sharma[1], Siddhartha Paudel[3], Abeer Alsadoon[4,5], Madhav Parsad Koirala[2] & Sumeet Gill[6]

This primary research paper emphasizes cross-validation, where data samples are reshuffled in each iteration to form randomized subsets divided into n folds. This method improves model performance and achieves higher accuracy than the baseline model. The novelty lies in the data preparation process, where numerical features were imputed using the mean, categorical features were imputed using chi-square methods, and normalization was applied. This research study involves transforming the original datasets and comparative model analysis of four Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) cross-validation methodologies to heart disease open datasets. The objective is to easily identify the average accuracy of model predictions and subsequently make recommendations for model selection based on data preprocessing cross-validation model increased (5 to 14%) more than baseline model for best model selection. From comparing each model's accuracy scores, it is found that the logistic regression and k-nearest neighbor models achieved the highest accuracy of 81% among the four models when single accuracy is a concern. However, the random forest model summary statistics attained an F1 score of 95%, precision (96%), and recall (97%), indicating the highest overall macro accuracy score. These findings can be further compared using learning curve validation. Conversely, the logistic regression model exhibited the lowest accuracy of 84% among the four machine learning models. However, this research does not cover hyperparameter optimization, which could potentially improve model performance.

Keywords Machine learning, Cross-validation, Accuracy-precision, Learning curve, Health informatics

Healthcare has been transformed by machine learning, which holds promise for better diagnostic precision in complicated illnesses like heart disease. However, optimal validation techniques for robust heart disease prediction models remain underexplored, highlighting the knowledge gap by enhancing diagnostic accuracy, particularly in complex conditions like heart disease causes. As heart disease continues to be a leading cause of death globally, the need for precise diagnostic tools has never been greater. However, optimal validation techniques for robust heart disease prediction models remain underexplored. Machine learning is the process of designing a model based on training and testing data sets whose value is further evaluated from validation sets of the sample. The train-test split is a widely used method for dividing research datasets into training and testing subsets. Model accuracy primarily depends on the input and validation sets. Cross-validation, with multiple-fold iterations, helps fine-tune the model to achieve optimal performance scores. Machine learning models typically begin by splitting the available dataset into training, validation, and testing sets, often using a ratio of 70:15:15. The model is built and trained using the training data, evaluated on the validation set to fine-tune its performance, and finally tested on the unseen testing set to assess its generalization capability[1]. Multi-grade classification and prediction based on previously scored grade patterns are more accurate when doctors use the medicine on asking some critical question solutions during examination. However, this approach may be less effective when

[1]IIS (Deemed to be University), Jaipur, India. [2]Pokhara University, Pokhara, Nepal. [3]IOE, Pulchowk Campus, Patan, Nepal. [4]Western Sydney University (WSU), Sydney, Australia. [5]Asia Pacific International College (APIC), Sydney, Australia. [6]Maharshi Dayanand University, Rohtak, India. ✉email: rimal.yagya@gmail.com

doctors are prescribed medicine without prior exposure to critical question papers[2]. Cross-validation is a method of training a model by dividing the dataset into multiple folds, using a portion of the data from each split as the validation set while the remaining data is used to train the model, ensuring optimal accuracy[3]. The most common stratified class validation is used to split the data, spreading a similar ratio from target outputs between prediction samples that provides the best average score[4]. The hold-out method works on the left part of the training sets for the model, while the stratified n-fold works on imbalanced data sets when each fold contains appropriately the same strata of each output class. The leave-out cross-validation uses samples to train and points as validation, which is repeated for all combinations, and the error is averaged until there is no randomness with averaging[5,6]. Logistic regression, random forest, support vector machine, bootstrapping, and cross-validation techniques are used to solve overfitting problems in medical research. Bootstrapping uses minimal sample data to resample the data, while cross-validation techniques use many contributing features to compare target responses[7]. Author[8] split the dataset into folds, trained the model on the training set, and validated it on the test set. Repeat these steps 3 to 6,000 times, with the first convolution reserved for model testing and the rest used for model training. Bias measures the difference between the model's predictions and the actual target values, while variance reflects the inconsistency of the model's predictions across different datasets. Ideally, the model achieves a balance between bias and variance, ensuring optimal explaining ability and yielding the best overall performance on the dataset[9]. Similarly authors[10] utilized the generalization process to evaluate how effectively a model is trained to identify meaningful data patterns and classify unseen data samples. Overfitting models remember the data patterns of the training dataset but do not generalize to unseen data, leading to high variance[11]. Underfitting occurs when the model fails to capture patterns from the dataset, often due to insufficient or low-quality training data. Additionally, a lack of adequate training samples can also lead to underfitting. Researchers aim to achieve a good fit by identifying patterns in the training data, which depends on the quality of data inputs and the splitting of folds during model development. Similarly, the author[12] proposes a three-phase model based on artificial neural networks (forward, backward). This analysis model showed and achieved a classification accuracy of 88.89% using the university dataset. If the neural network backpropagation model showed 85% accuracy when testing records of unseen data. Similarly[13], conducted research on nonsmoking among children aged 12 to 19 years, reporting accuracies ranging from 76% (minimum) to 94% (maximum) variations. Using assembly physical movement, body mass list, and blood glucose did not progress; the predominance declined from 70 to 60% over the same time[14]. The classification framework and accomplished framework show 89.1% accuracy, however, model wise differs largely by 80.09–95.91% individually utilizing ventricular systolic execution within the distribution the distributed reports shift broadly from 13 to 74%; the detailed yearly mortality rate moreover shifts from 1.3 to 17.5%. Similarly, authors[15] integrated medical decisions based on cardiac infection symptom framework classifiers, multi-layer perceptron, artificial neural network-driven methodology highlights for heart disease, machine learning algorithms, artificial neural networks, and artificial neural networks using analytical hierarchical fuzzy processing. Their proposed classification system achieved a classification accuracy of 91.10%. This work mainly discusses the model selection and its accuracy, but without dealing with various cases of overfitting and underfitting classification computation double classification problems, y [0, 1], negative history, and an estimate of the forward variable y for one positive course (AUCROC). The multiclass to predict estimates of y for y [0, 1, 2, 3] increased variance. A guess is sketched to classify 2 classes and 1 class; the yield of the classifier is a 0.5 threshold value. A support vector machine can be a machine learning classification computation commonly used for classification problems. The support vector machine used the most extreme edge technology modified[16] of image analysis. According to[17], imbalanced data sets classified primary school and higher education using the multiple-choice online of Bharathiar University. The research conducted on the number of students not completing graduation in the USA is 20%, and in Europe, it is around 20–50% to finish their studies on time that utilizing multigrade target features. Similarly, the authors[18] of relevant studies published between 2000 and 2018 indicate that multiple factors influence performance in non-linear ways in online learning of performance analysis and influence factor identification based on the behavior of assessments, teaching, and association rule mining. The regression, classification analysis, and performance prediction of a majority of 46% of modeling studies prefer to classify the performance as success or failure too. Similarly, author[19] used supervised mastering algorithms for the improvement of a predictive version of the Federal Board of Intermediate and Secondary Schooling Islamabad Pakistan, the use of folding 10. In k-fold pass-validation, a reduced education vector-based totally aid vector device capable of predicting at-risk and marginal college students that the support vector completed a training vector discount of at least fifty-nine point 7% without changing the margin or accuracy of the classifier. Moreover, the effects confirmed the proposed approach to be able to achieve a basic accuracy of 91% and 93.8% in predicting at-risk, respectively, which are good examples of standalone accuracy variation. Similarly, the authors[20] used light gradient boosting, extended gradient boosting, random forests, and multilayer perception classifiers from UCI records and used three groups for error prediction with stack generalization about machine learning repositories to target features. They achieve an average sensitivity of 97.3%, joint accuracy of 97.2% classification, an F1 rating of 97.1%, and an average (98.86%) neural network algorithm. The drop rate down from 12 to 1.14%. Similarly, author[21] used a neural community version for the exhibit that the proposed model reaches up to an accuracy of 95%, which is higher than many present methods for cerebral infarction disorder. Similarly, authors[22] of this DBSCAN identified and extracted nine clusters of informative gene facts selected by differential gene expression analysis to five special category models. Then, a deep mastering method is hired to ensemble the outputs of the 5 classifiers. Similarly, authors[23] The modified J48 classifier is used to boost the accuracy fee of the data mining technique. The facts mining tool MATLAB for generating the decision classifiers and Naive Bayesian classifiers in WEKA. The general accuracy is around eighty-three. Similarly, authors[24] memetic set of rules stepped forward the accuracy from 88.0% to ninety-three point 2%. Which additionally found that the memetic algorithm had a higher accuracy than the version from the genetic set of rules and a regression model. Similarly[25], uses a genetic algorithm-primarily based regression model for predicting inflation levels. The version becomes educated and evaluates the usage of facts. Similarly, the authors[26,27] used a prediction model delivered with one-of-a-kind combos of features and several acknowledged category strategies. That produces an improved performance level with an accuracy degree of 88.7% via the prediction version for coronary heart disorder with the hybrid random forest area with a linear model. Similarly, authors[28] used the prediction of the performance of the k-nearest neighbor algorithm for the overall performance class, and overall accuracy of the tested classifiers is acquired at 60%. The decision tree categorized approximately 72.51% for the 10-fold go-validation testing and sixty-nine 66% for the share split testing. The precision is high for the primary class (67–76%) and 2d elegance (72–85%). Similarly, authors[29] did research to locate which function has the best effect on the goal class to locate which method outperforms the most used RF component, J48, Bayes internet Socio-financial, demographic, and educational facts random forest provided 90% accuracies. The internal assessment influences the final semester percentage. Similarly, author[20] used the demographics to outperform random forest by providing 99.90% accuracy on training information with 10-fold cross-validation and 99.82% accuracy on the holdout method. While the guidance is implemented, the accuracy of the simplest ANN improves by up to 100%, and vice versa for other methods. Self-efficacy and motivation for success are good predictors while correlating with GPA. Like demographic features, heart disease independent features were significantly influenced by target heart disease, and besides models, it also varies their accuracy, so this research tries to validate cross-validation scores, including four popular model comparisons.

Methodology

Machine learning plays a giant position in extracting the hidden capabilities from the scientific records, beneficial for early detection from the heart ailment report repository; that's the reason approximately 12 million deaths happen globally[30,31]. Coronary disorder dying is observed greater in the USA than in other advanced European countries[32]. Therefore, this research aims to achieve the most accurate validation accuracy prediction for machine learning models using four techniques: logistic regression, support vector machines, nearest neighbors, and random forests. The validation focuses on identifying coronary artery conditions, making it easier for medical practitioners to select the most suitable model for classification and prediction. This study compares four machine learning models—logistic regression, support vector machines, k-nearest neighbors, and random forests—using cross-validation to determine their predictive efficacy.

Data preparation flow chart

The datasets on heart disease comprise 13 categorical attributes requiring preprocessing before conducting machine learning model testing and evaluation, depicted in Fig. 1. After loading the dataset into the Python console, the Python command df. Types describe the data types categories with their respective categories. Each feature with a unique c command describes the properties of the data sets. One hot encoder (categories='auto') fits with fit transform (df [sex, cp., fbs, restecg, exang, slope, thal, ca]). The column names of each categorical OneHotEncoder constitute 76 columns of data sets. Cross-validation provides better model optimization of heart disease using logistic regression and support vector machine learning models before finalizing the best model for research data. A dataset of df = pd. read csv ('https://raw.githubusercontent.com/kb22/Heart-Disease-Prediction/master/dataset.csv') was used. The most important details in this text are the age, chest pain, treetops, chol, thalach, oldpeak, m, f, typical angina, atypical angina, non-anginal pain, asymptomatic, normal, abnormal, normal, abnormal, yes, no, upsloping, flat, down, normal, fixed defect, reversible defect, non, Ca0, Ca1, Ca2, Ca3, and Ca4's[33]. Similarly, author[33]. Similarly, author[34] used a heart disease dataset of samples with 14 independent samples and final target variables having heart disease 1 and not having heart disease 0 as target variables, which

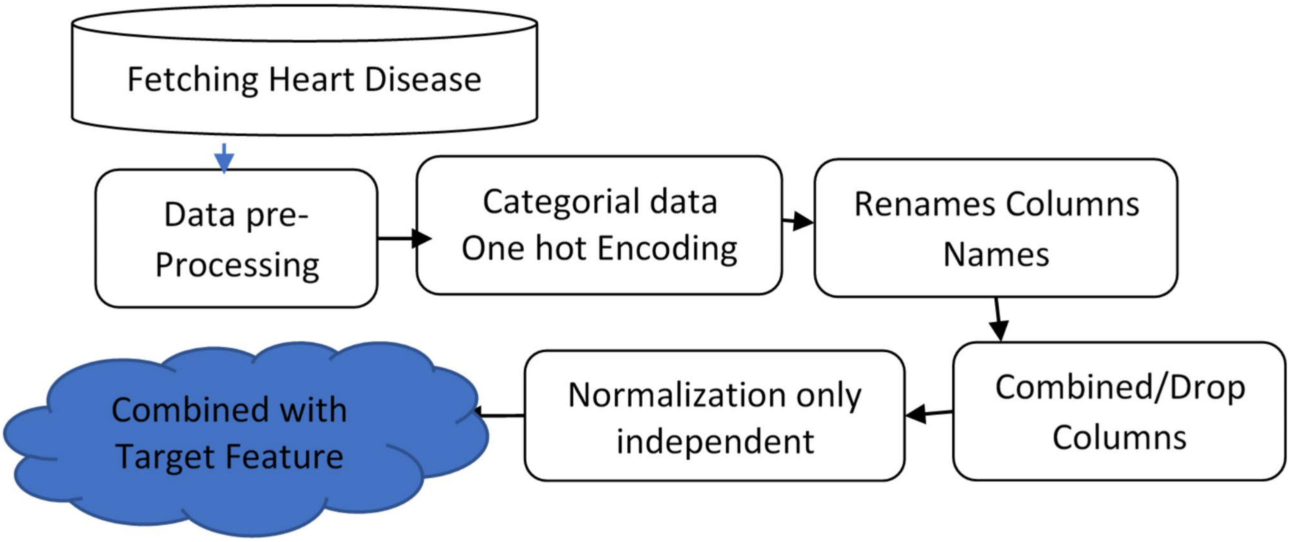
Fetching Heart Disease
Data pre-Processing
Categorial data One hot Encoding
Renames Columns Names
Combined/Drop Columns
Normalization only independent
Combined with Target Feature

Fig. 1. Steps in data preprocessing for the heart disease dataset.

Age	trestbps	chol	thalach	oldpeak		Age	trestbps	chol	thalach	Oldpeak
63	145	233	150	2.3	Before After	0.952	0.763	-0.256	0.0154	1.08
37	130	250	187	3.5		-1.91	− 0.092	0.0721	1.633	2.122
41	130	204	172	1.4		-1.47	-0.092	− 0.816	0.977	0.3109
56	120	236	178	0.8		0.18	-0.663	− 0.198	1.239	-0.206

Table 1. Data table Preparation before/ after.

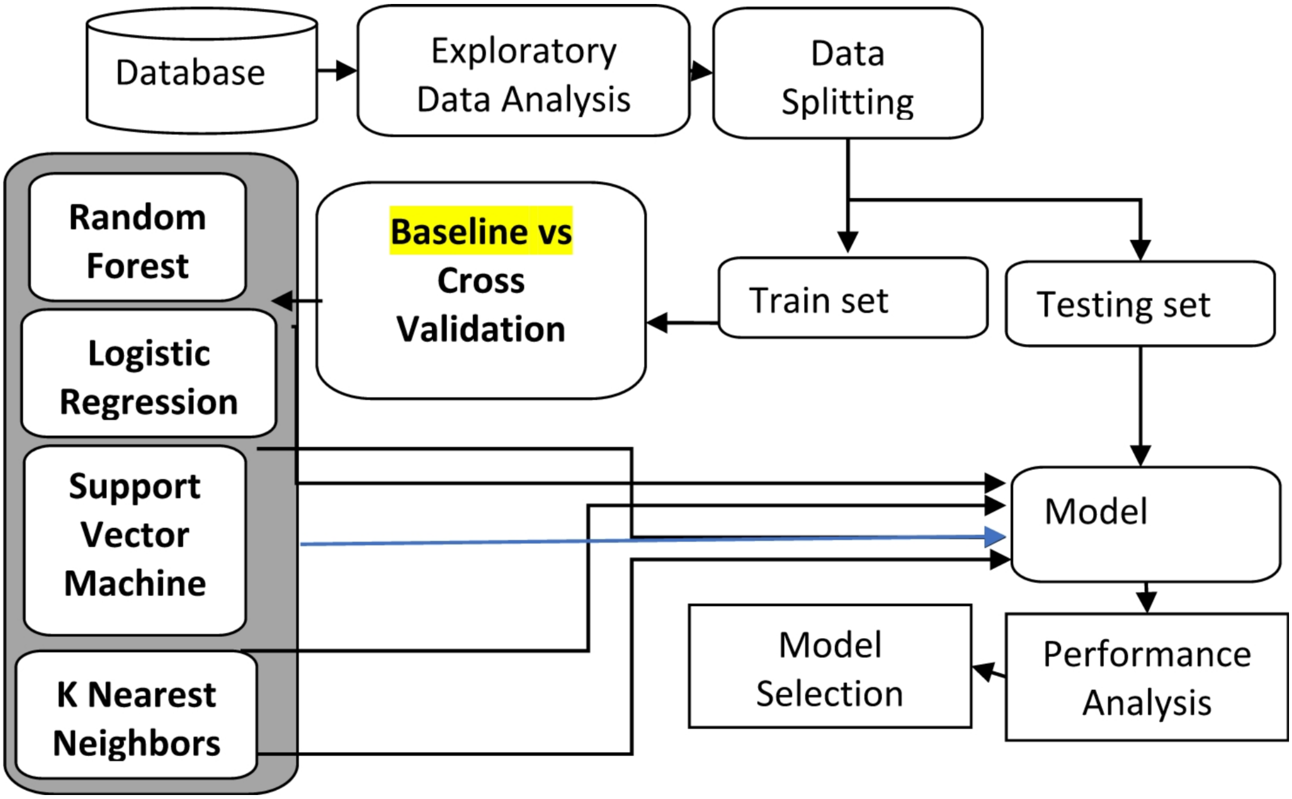

Fig. 2. Validation design diagram.

had 303 records of heart disease patients viewed and downloaded 381,647 and 62,705, respectively, till January 2023. After loading the sklearn preprocessing library of the standard scaler into the Python console, rename their respective columns as final2[age, trestbps, chol, thalach, oldpeak, m, f, and normal][35].

After being combined with the target column and normalized, the dataset is ready for model comparison, as illustrated in Table 1. The machine learning model always needs train-test splits. This research forest used four model comparisons using normalized data with 80:20 splits, and the parameters using stratify indicate that in each of these datasets, the target/label data proportion is preserved as 50:50 for the classes [0, 1]. This indicates there would not have to be oversampling and under-sampling problems of both training and test sets. And with the random state = 42, we get the identical teach and test units across different executions, but this time, the teach and check sets are exclusive from the preceding case with random state = 0. The train and test units immediately affect the model's performance score.

Validation design diagram

This study aims to validate the most effective machine learning model by employing four widely recognized algorithms, as depicted in Fig. 2. This study compared baseline results with cross-validation using normalized datasets. The datasets were uniquely resampled with various n-splits and tested across machine-learning models, including logistic regression, random forests, support vector machines, and k-nearest neighbors. The process begins with preparing the datasets by applying one-hot encoding for categorical variables and normalizing the data. This is followed by predicting individual model accuracies and performing a 5-fold cross-validation after splitting the preprocessed heart disease datasets into training and testing sets. This output will further plot using a learning curve with the same cross-validation. The confusion matrix and ROC-AUC curve for each model were compared alongside their respective model summaries, incorporating the learning curve at each step. The study adopts a rigorous approach to data preparation and model validation. The dataset, consisting of 13 categorical attributes, undergoes preprocessing and feature engineering, where one-hot encoding expands the data to 76 columns. An 80:20 train-test split with stratification is used to maintain class proportions, followed by a 5-fold cross-validation procedure for robust evaluation. Additionally, a 5-fold cross-validation is employed to plot learning curves, ensuring a comprehensive assessment of model performance. Using Python's Seaborn for heatmaps. So, data scientists received the best model selection of similar data samples.

Results and discussion

After designing the data sets, the correlation between dependent and independent variables is described using a heatmap (final2.corr(), cmap='cool warm') in the case of displaying each correlation value. The sns. heatmap (final2.corr(), annot = True) function is used. Correlation plots are used to understand which variables are significantly related to each other and the strength of this relationship to cause heart disease.

The baseline correlation coefficients between the dependent and independent features of the heart disease dataset, derived from logistic regression analysis (Fig. 3), show a mean squared error for features such as sex, cp., true slope, fats, thalach, and ca., indicating satisfactory performance below 10. In contrast, independent features like exang (38), oldpeak (29), and slope (58) exhibit significantly higher error values. Similarly, the model's coefficient of determination for the dependent variable, as explained by the independent variables, demonstrates substantial variation.

Similarly, the researcher examines the relationship between the dependent and independent features using binary logistic regression to evaluate feature associations. The R-squared metric measures the proportion of variance in the dependent variable explained by the independent variables, as shown in Table 2. In this analysis,

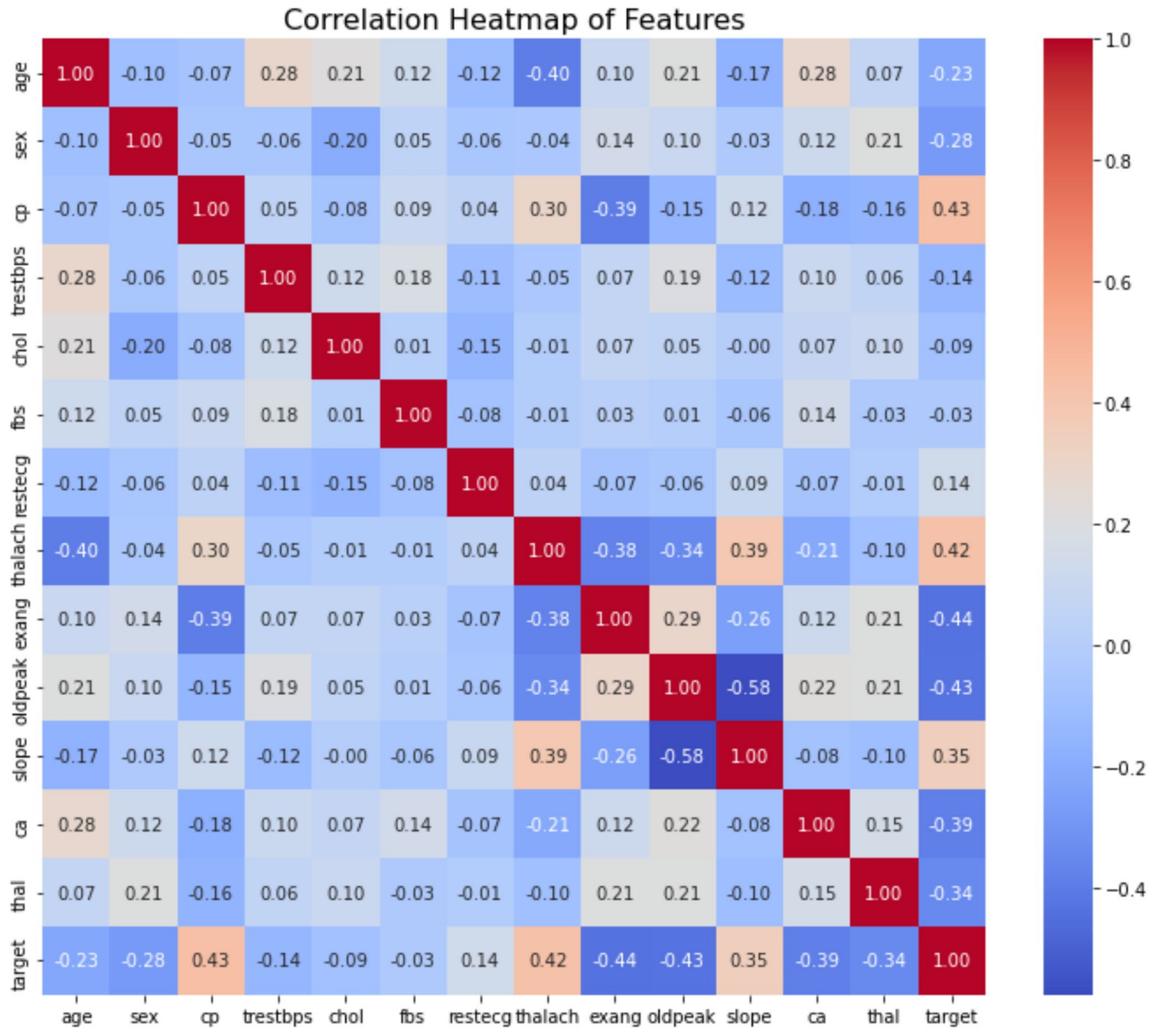

Fig. 3. Data correlation heat map.

58% of the variance is explained by the independent variables. The adjusted R-squared, which provides a more accurate measure, stands at 54%, which was satisfactory. The p-value (P) is close to zero, indicating strong evidence against the null hypothesis, indicating there is a significant contribution of dependent features. The F-statistic (F) is used to test the overall significance of the regression model. A higher F-statistic indicates a more significant relationship between the independent variables and the dependent variable, which is low, 17.71 in total. The intercept represents the expected value of the dependent variable when all independent variables are zero. In this case, the intercept is 0.54, with a high t-value and low p-value. The coefficients indicate that a one-unit increase in age is associated with a 20-unit increase in the dependent variable, but not statistically significant. The coefficients suggest that a one-unit increase in Trestbps is associated with a 41-unit decrease in the dependent variable. The Sex_1 variable is statistically significant, with a t-value of 2.64 and a low p-value (0.001). The other variables (age, trestbps, cholesterol, thalach, oldpeak) do not seem to have a significant relationship with the dependent variable, showing each feature significantly. Therefore, this research needs to further evaluate the machine learning model with cross-validation for the evaluation of prediction accuracy of heart disease patients.

Baseline machine learning model without cross validation

The default machine learning model of four different machine learning model parameters (models = [logistic regression with maximum iteration of 1000, SVC with kernel is linear, neighbors classifier, random forest]) was designed in the model, and then using a loop, the model whose accuracy score was calculated using after fitting the models. This process involved splitting the data into training and testing subsets, designing the model, making predictions on the test data, and calculating each model's accuracy score. The workflow consisted of fitting the model using a model fitted with train and train, then predicting the model's accuracy. The console output revealed that the logistic regression and nearest neighbors models achieved the highest accuracy (81.9%), followed by SVM with a linear kernel and scoring 78.6%. Specifically, logistic regression is 81.9%, K Neighbors is 81.9%, the kernel is linear at 78.6%, and Random Forest scored 78.6%. These accuracy scores demonstrate the models' performance using default settings. Additionally, the classification accuracy score in the multilabel problem was calculated based on the yttrian sample, where true labels matched predicted labels exactly. This measure effectively highlights the models' behavior in handling the given dataset Fig. 4: Confusion matrix of four models. The confusion matrix for each target class is generated using the plot of the confusion matrix function, which visualizes class-specific issues and evaluates the model's error patterns, as shown in Fig. 4. The rows represent the actual classes, while the columns indicate the predicted classes for a binary target. The off-diagonal elements highlight the incorrect predictions, making it easier to identify errors. The model's accuracy, evaluated using the classification report, was initially calculated under train-test splits of the entire dataset (x and y). However, due to significant variation in the results, cross-validation was employed. In this approach, test samples were varied across five iterations, ensuring a more robust evaluation. The summary statistics for all models, including classification reports, were printed using the command classification report with y, model. predict (x)).

Table 3 summarizes the maximum accuracy achieved by the models and highlights the minimum performance across all four algorithms. Logistic regression and K-nearest neighbor models demonstrate the highest predictive accuracy (81.9%) compared to support vector machine and random forest models. However, when evaluating macro-average metrics such as precision, recall, and F1-score, the Random Forest model outperforms, achieving an accuracy range of 93–97% for true matches. This suggests that while logistic regression is preferable when accuracy alone is the primary concern, the random forest model is more suitable for scenarios where overall performance, including precision, recall, and F1-score, is critical for decision-making.

Machine learning model using cross-validation

Another way of resampling heart disease data sets for machine learning is using cross-validation techniques. Therefore, cross-validation is a process for evaluating all K-fold sample models by training each training/test model on subsets of the data. The final majority of the vote will be predicated after evaluating them on the complementary subset of the data. This process is quite good while designing the cross-validation to detect overfitting problems to generalize a pattern. Each model was individually evaluated and fitted, and their accuracy score was calculated using cross-validation and accuracy functions.

R-squared:0.58	Adjusted *R*-squared:0.54		*P*: 5.38e-41	F-stasticts:17.71
	coef	Std	t	P
Const	0.54	0.019	28.29	0.00
Age	0.20	0.024	1.023	0.30
Trestbps	-0.41	0.021	-0.92	0.56
Chol	-0.016	0.026	1.67	0.42
Thalach	0.042	0.027	-1.80	0.9
Oldpeak	-0.041	0.022	-3.40	0.07
Sex_1	-0.07	0.020	2.64	0.001
Fbs_1	0.011	0.043	0.58	0.00

Table 2. Logistic summary statists.

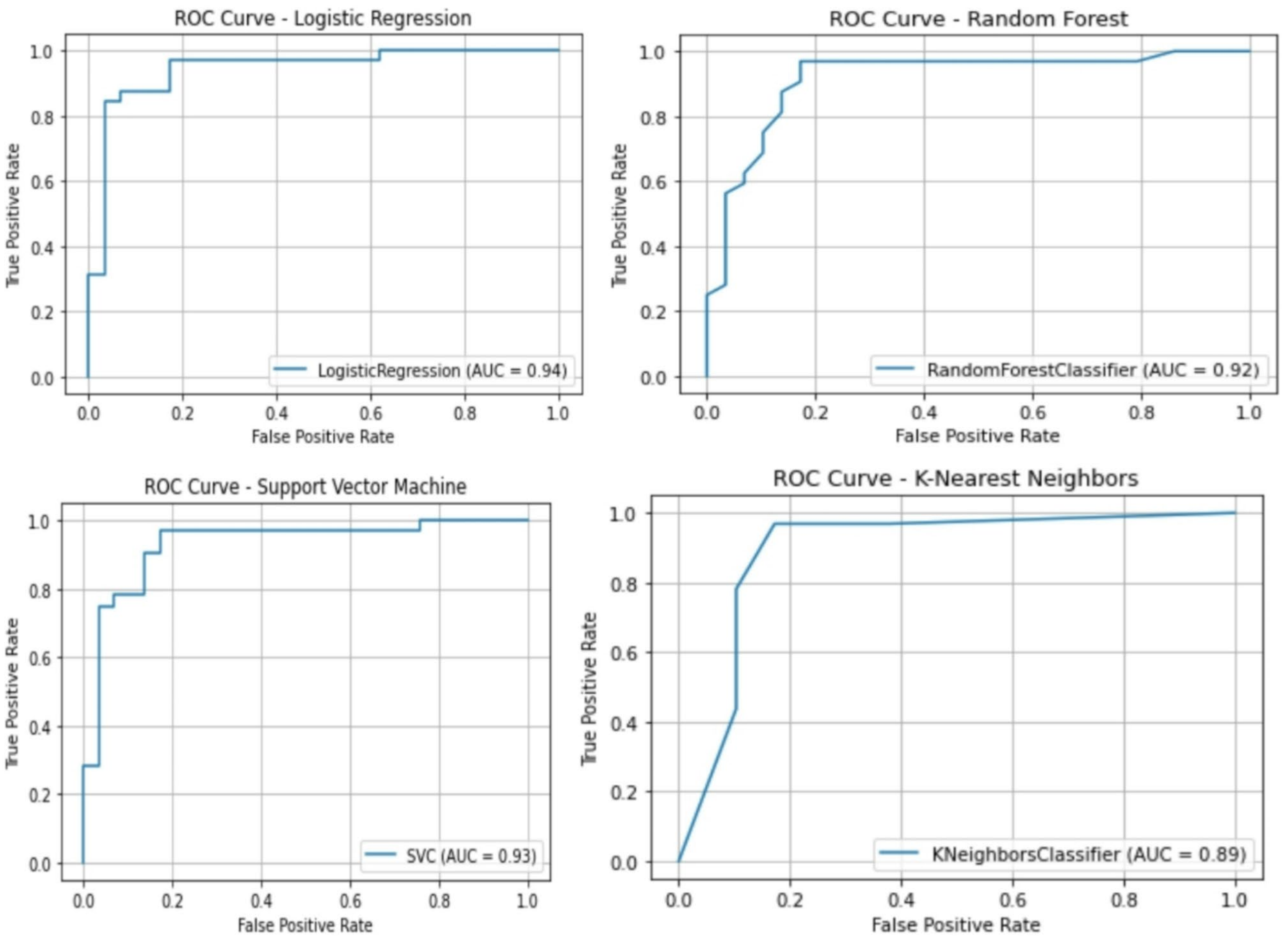

Fig. 4. ROC/AUC curve of each model.

	Minimum				Maximum		
	Accuracy %	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Logistic regression	81.9	86	88	87	86	83	84
SVC classifier	78.6	88	91	88	85	81	85
K NN classifier	81.9	89	92	89	86	83	86
RF classifier	78.6	96	97	96	95	93	95

Table 3. Maximum and maximum accuracy.

Model/iteration	1	2	3	4	5	Average
Logistic regression	88	88	80	83	78	83.81
Support vector	88	88	75	81	78	82.49
K-nearest neighbor	85	86	81	85	81	84.15
Random forest	83	90	80	85	81	84.15

Table 4. Machine learning model accuracy score using cross-validation.

Table 4 outlines the average accuracy obtained through 5-fold cross-validation. The random forest and k-nearest neighbor models achieved the highest accuracy at 84.15% because the model can accept both data sets by applying decision tree separation in tree-like structures. Following closely, logistic regression attained an accuracy of 83.81%, while the support vector machine model achieved 82.49% accuracy with the normalized data sample. Notably, the random forest algorithm recorded a maximum accuracy of 90% in certain instances. Using the upgraded data sample, a 5-fold cross-validation process was performed to evaluate and compare

Model/iteration	1	2	3	4	5	Average %
Logistic regression	88.5	88	80	83	78	83.81
Support vector	88.5	88	75	81	78	82.49
K-nearest neighbor	85.2	86	81	85	81	84.15
Random forest	86.8	86	78	86	87	84.73

Table 5. Machine learning model accuracy using combined validation.

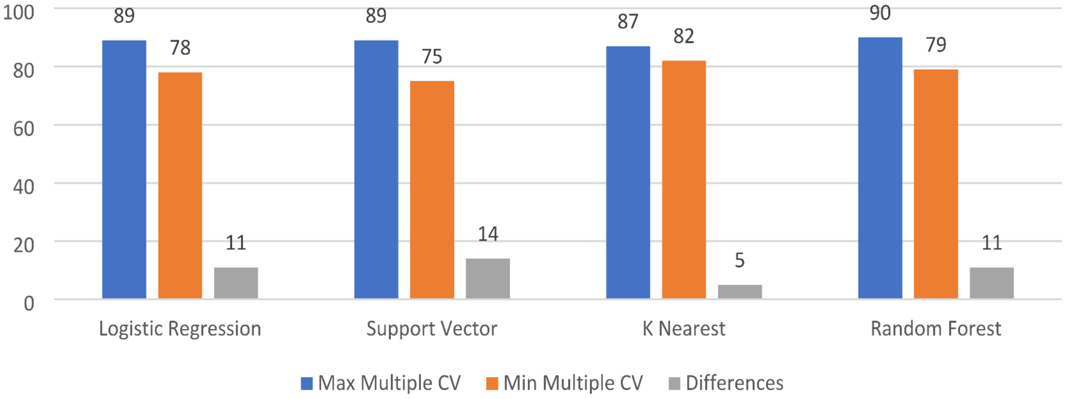

Fig. 5. Min-Max bar chart accuracy comparison.

machine learning models, including logistic regression with a maximum iteration equal to 1000, SVM with kernel='linear, neighbors, and random forest algorithms. For each model, the cross-validation scores were computed using a cross-validation score at 5-fold. The mean accuracy was calculated as the average of the cross-validation scores, expressed as a percentage at two digits. The resulting table highlights the accuracy for each fold and the average accuracy of the sample data.

In Table 5 above, the combined machine-learning models with various hyperparameters display the accuracy scores for each model. The k-nearest neighbor model yields the highest accuracy at 84.15%, followed by logistic regression and random forest models with the second lowest accuracy at 83.83%. The support vector machine model exhibits the lowest accuracy, scoring at 82.2%. Therefore, a researcher might take either the highest score or the lowest score for evaluating the model accuracy for heart disease. The model might be confused due to taking max/min from the five cross-validated accuracies. The accuracy score using the max/min of each model return value depends on the setting for the normalized parameter due to the sample reshuffled using the stratified value becoming true when the researcher considered the sample reshuffled when the cross-validation iteration; the difference between each model matters for large model accuracy for correctly classified samples.

From Fig. 4, the best-tuned models tested with prediction values and inversely tuned parameters achieved the following accuracy scores: Logistic Regression (94%), Random Forest (92%), Support Vector (93%), and K-Nearest Neighbors (slightly lower). All models demonstrated accuracy scores exceeding 90%, indicating satisfactory classification performance when combined and tested.

In Fig. 5 above, a bar diagram compares the accuracy scores of each model with the highest and lowest scores, both with and without loop accuracy considerations. The logistic regression model and support vector model made a large difference between without cross-validation and with cross-validation. The support vector machine had the largest difference, whereas the nearest neighbor's model produced the best result among the four models. Similarly, when comparing a single independent model vs. multiple with validation model support vector, it differs by 14% as compared to k. Nearest model accuracy is below 5%, as depicted in Fig. 6. The baseline accuracy of a machine learning model becomes lesser than cross-validation due to standard reshuffling with n-fold while model tuning.

Similarly, it is concluded that the individual model has the least accuracy with mean values. Therefore, it is recommended that the max-multiple cross-validation model produce the highest accuracy. The random forest model with CV scored 90% accuracy as compared to 75%. Random Forest performed better due to its ensemble nature, effectively reducing variance and handling imbalanced data. Cross-validation compares model performance by splitting data into training and testing sets multiple times.

Learning curve of all models

The learning curve represents the computational cost and effort involved in acquiring knowledge or skills over time or through repeated experiences. Learning curves visualize the challenges associated with mastering a subject over a given period and the relative progress made during the learning process. The concept assumes a doubling of output, where, for example, a 70% learning curve indicates that the cumulative average time per unit decreases to 70% of the previous average as the output doubles. This is measured from the first unit produced.

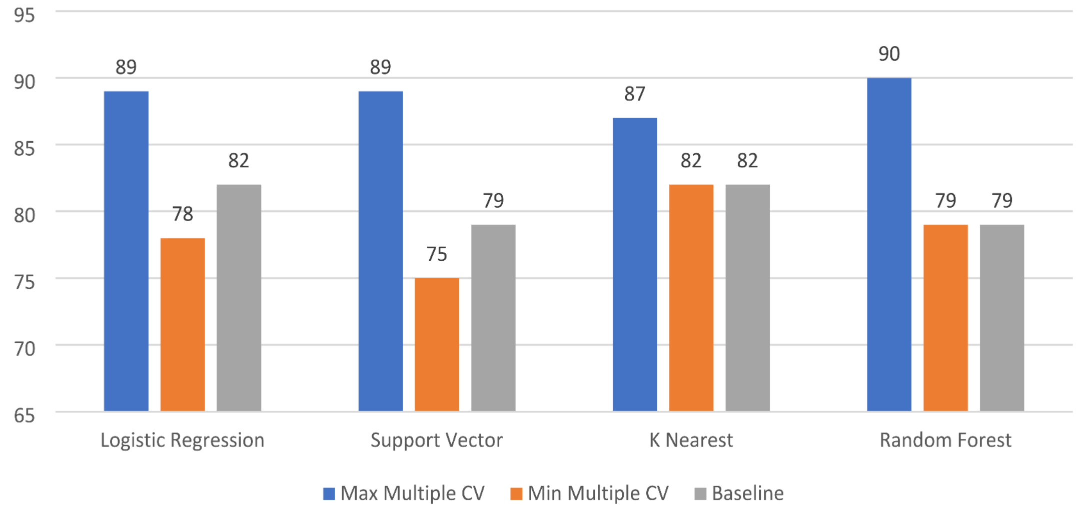

Fig. 6. Ensemble model bar chart comparison.

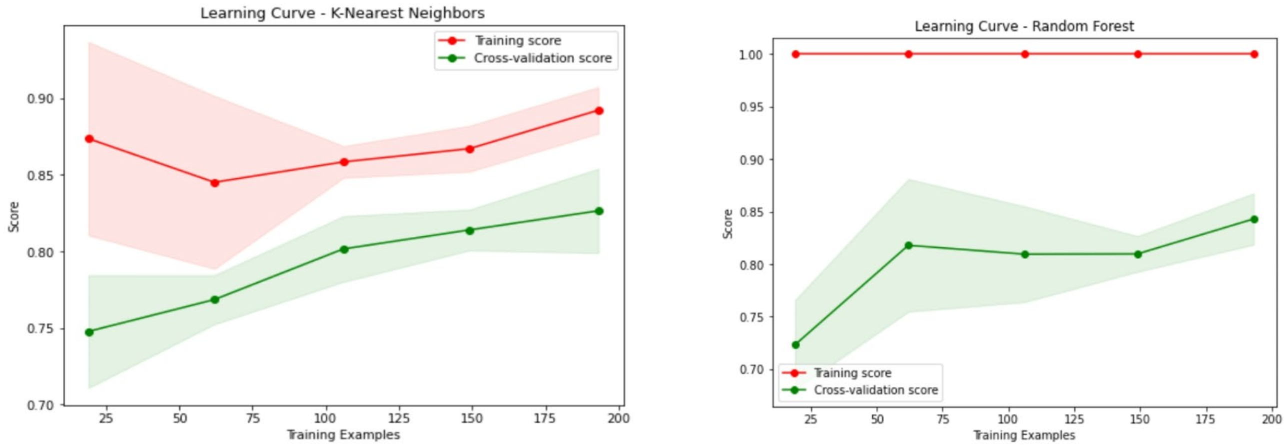

Fig. 7. Learning curve k-nearest (**a**) and Random Forest (**b**).

Learning curves are often generated using methods like grid search cross-validation, which evaluates models or functions across a range of parameters, leveraging absolute numbers of training examples to plot and assess performance trends. The scoring is used as an evaluating metric for the version performance to determine the fine hyperparameters; if not special, then it uses an estimator rating. Through default, it is ready as five; however, here the researcher decided on 10 reputations. The jobs symbolize the wide variety of jobs to be run in parallel, and – 1 indicates the application of all processors. After importing the learning curve package in the Python console, the normalized data first splits into a dependent a and an independent set of heart disease, X = final4. drop([non, target], axis = 1), and y = final4 with the target. The learning rate splits with scoring accuracy, and the learning rate starts from 0.01, 1, 50, and 100 iteration splits. After the train test splits, calculate the means of accuracy of the K-nearest model plot. The learning curve describes the training and validation metric for describing overfitting and underfitting.

In Fig. 7 above, the two lines represent the validation curve, which changes gradually, with the lower line indicating the training error or accuracy score. This curve describes how the error metrics when increasing training and validation of the model best fit. Each line describes the combined effects of the model with heart data sets. Initially, when the model reached up to 100 training sets, the model produced a straining line, and then after both were reduced for conversion, reaching 250 iterations produced constant output due to heart disease data sets indicating high variance. Similarly, when the Random Forest learning curve was produced after 50 training samples, the machine learning mode increased the high accuracy score (84%) but indicated high bias as compared to the line.

In Fig. 8 above, the maximum variation is depicted as each cross-validation iteration changes during the model training and testing process. The support vector machine and logistic regression study the use of heart sickness data set step by step to validate after 50 generations. The training facts curve shows greater gradual improvement at 200 steps hastily than the validation curve, which in the end suggests overfitting. The system

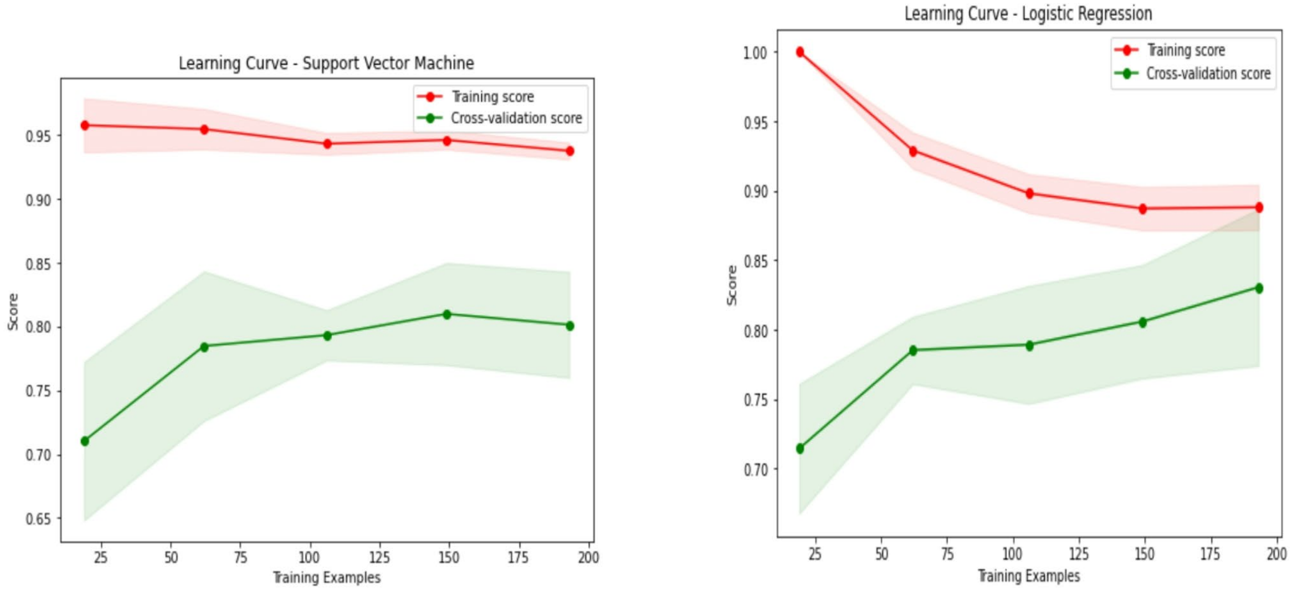

Fig. 8. Learning curve of support vector (**a**) and Logistic regression (**b**).

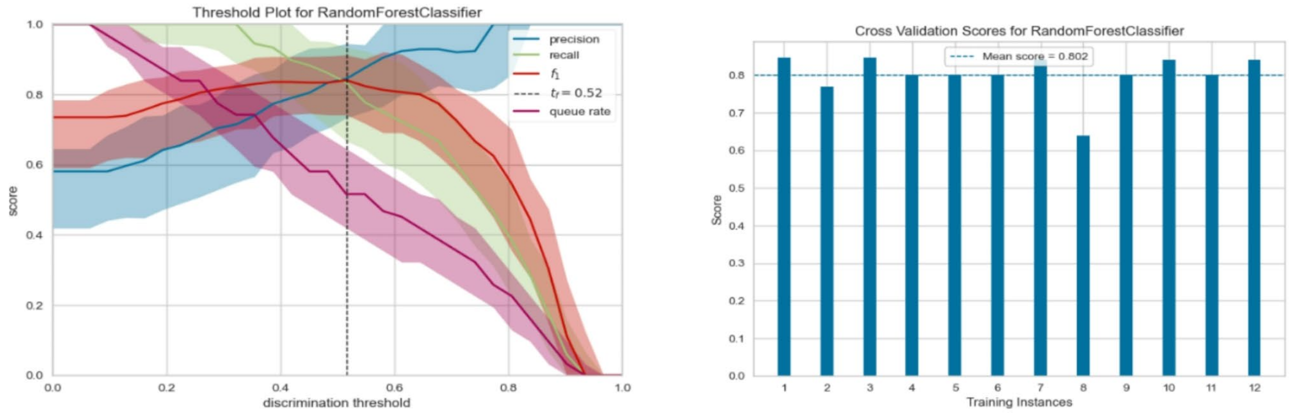

Fig. 9. Accuracy score and threshold (**a**) Accuracy at 12 iteration (**b**).

gets to know the curve is useful for many purposes, together with evaluating distinct algorithms, choosing model parameters for the duration of layout, and determining the number of statistics used for training. This variance in the dating between practice and proficiency over time is referred to as the ‘mastering curve. The logistic regression tuned more than 85% best tuned when both crossed fitted after 175 iterations.

The dataset, consisting of 303 records, is further divided into 165 and 138 records for testing purposes. Discrimination threshold plots with 100 trials showcase precision, recall, and F1 score plots with both training and testing unseen datasets, revealing the best fit at 84%. Cross-validation scores might vary within ± 50, as illustrated in Fig. 9(a). Similarly, accuracy scores from 12 iterations show a mean squared score of 80.2%, yielding similar results, as shown in Fig. 9(b). After using random forest error cross-validation curves, 85% with 16 features were scored best optimal when 5 features were folded in each step.

Likewise, in Fig. 10(a), the prediction error and residual error reach 85.5 when the number of features becomes 16, as observed in the support vector machine model. Additionally, both the training and test samples exhibit symmetric histograms, indicating a uniform distribution for predicting values, as depicted in Fig. 10(b) and (c).

Following the separation of dependent and independent features, the logistic regression model in Fig. 11(a) reveals feature importance and confusion matrix analysis. The accuracy score reaches 88%, while the predicted AUC score attains 95%, signifying high precision. The feature importance of logistic regression shows both positive and negative contributions to target heart disease detection. Likewise, the random forest model shown in Fig. 11(b) achieves the second-highest accuracy at 93%, with a commendable AUC score. The macro accuracy for heart disease prediction reaches 94%, and precision, recall, and F1 scores are plotted accordingly. However, the k-nearest neighbor model predicts with comparatively lower accuracy among the models analyzed.

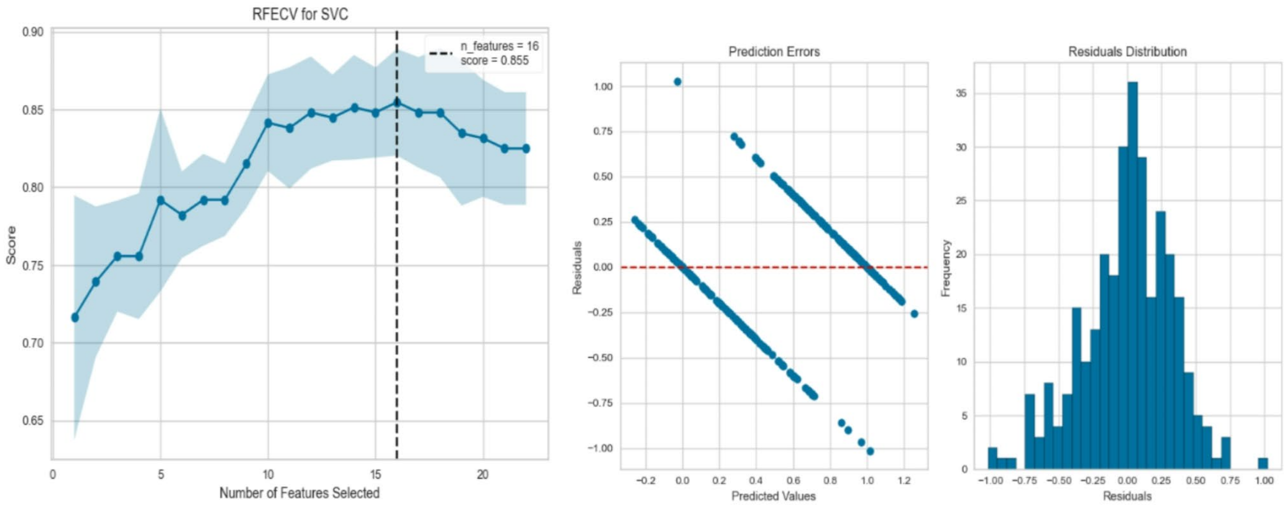

Fig. 10. Predication error (**a**) and residual plot (**b**) histogram (**c**).

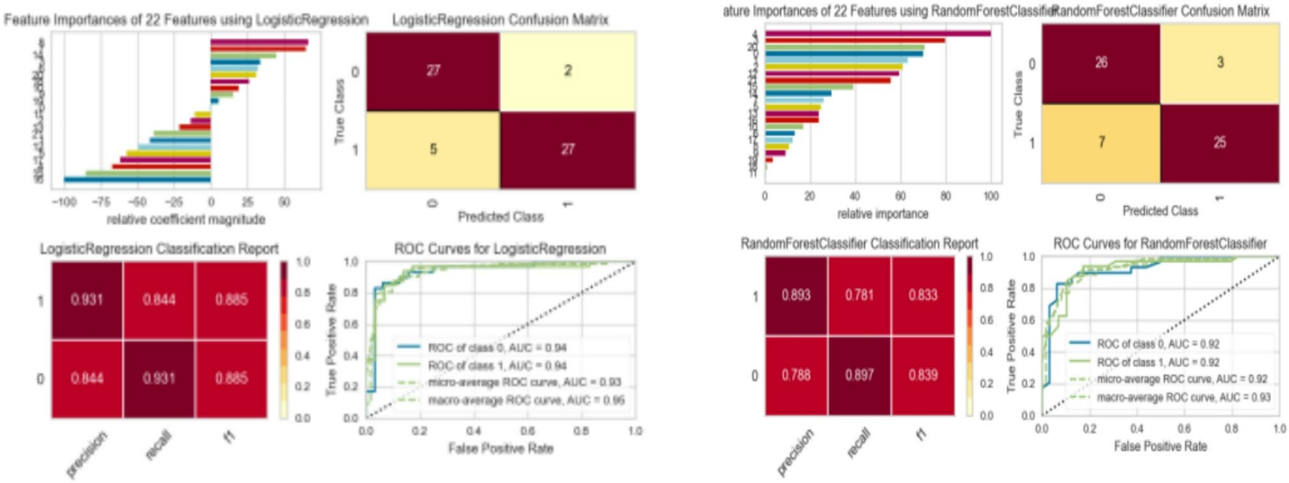

Fig. 11. Logistic regression (**a**) Random Forest summary statists (**b**).

In Fig. 12(a), the Support Vector Machine (SVM) model achieved an accuracy of 86% across four testing levels, with an Area Under the Curve (AUC) of 92% for predicting the absence of heart disease. Additionally, the macro average accuracy improved to 93%. Similarly, Fig. 12(b) highlights the performance of the K-Nearest Neighbors (KNN) model, which achieved a macro average accuracy of 93%. Its prediction accuracy was 89% for detecting heart disease cases and 81% for identifying non-disease cases. The superior performance of the Random Forest model further underscores its potential as a dependable tool for early heart disease detection, particularly in resource-constrained environments. However, the successful implementation of machine learning in healthcare must address critical concerns such as bias, interpretability, and data privacy to ensure equitable and effective outcomes.

Conclusion

Cross-validation is a robust statistical approach to evaluating machine learning models by systematically splitting data into training and testing subsets. This method ensures models generalize effectively to unseen data, minimizing the risks of overfitting and underfitting. Among multiple models tested, the Random Forest (RF) model achieves the highest macro accuracy (94%) and precision (97%), outperforming Logistic Regression (81%), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). SVM demonstrates 89% accuracy in identifying heart disease cases and 81% in distinguishing non-disease cases, while cross-validation highlights a 14% accuracy variation for SVM and less than 5% for KNN. These insights emphasize the importance of cross-validation in improving model accuracy and reliability. Learning curves further aid in understanding how models optimize parameters over time, with higher scores reflecting better performance. The baseline model without normalization achieved a cross-validation accuracy variation of ±14%, indicating potential

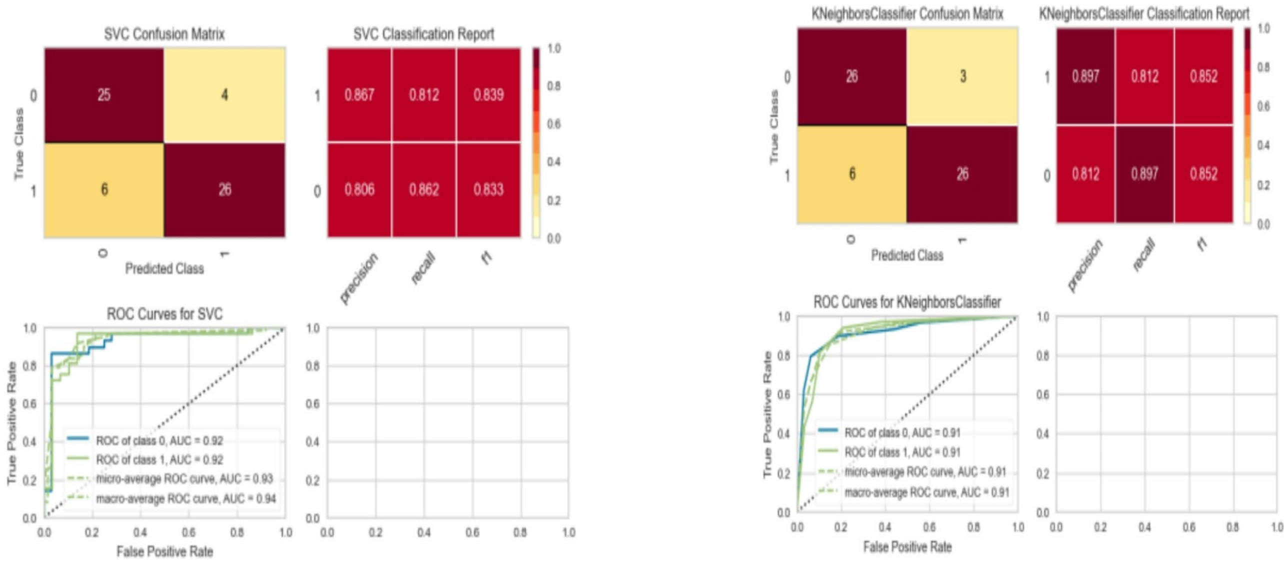

Fig. 12. Support vector (**a**) and K nearest summary statists (**b**).

improvements through advanced feature selection and ensemble methods. The integration of machine learning into healthcare offers immense potential for predictive accuracy and decision support. However, challenges such as bias, interpretability, and data privacy must be addressed to ensure equitable and reliable application in clinical settings. Future recommendations include refining feature selection techniques, leveraging ensemble models, and incorporating larger, real-world datasets to enhance model robustness and generalizability.

Data availability

The open-source dataset on heart disease, containing 13 features, is freely accessible at the following link: 'https://raw.githubusercontent.com/kb22/Heart-Disease-Prediction/master/dataset.csv') dataset. Additionally, the Python source code for migrating the source data to research data is openly available in my GitHub repository: https://github.com/yagyarimal/Heart22.

Received: 21 April 2024; Accepted: 10 March 2025
Published online: 18 April 2025

References

1. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R Stat. Soc. Ser. B Methodol.* **36**(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x (1974).
2. Chin, C. & Osborne, J. Students' questions: a potential resource for teaching and learning science. *Stud. Sci. Educ.* **44**(1), 1–39. https://doi.org/10.1080/03057260701828101 (2008).
3. Maldonado, S., López, J. & Iturriaga, A. Out-of-time cross-validation strategies for classification in the presence of dataset shift. *Appl. Intell.* **52**(5), 5770–5783. https://doi.org/10.1007/s10489-021-02735-2 (2022).
4. Mahesh, T. R., Geman, O., Margala, M. & Guduri, M. The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthc. Anal.* **4**, 100247. https://doi.org/10.1016/j.health.2023.100247 (2023).
5. Barrow, D. K. & Crone, S. F. Cross-validation aggregation for combining autoregressive neural network forecasts, vol. 32, no. 4. 1120–1137 (Accessed 14 Jan 2025). https://www.sciencedirect.com/science/article/pii/S0169207016300188 https://doi.org/10.1016/j.ijforecast.2015.12.011 (Elsevier, 2016).
6. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**(1), 1–36. https://doi.org/10.1038/s41524-019-0221-0 (2019).
7. Ye, Z. et al. Predicting beneficial effects of Atomoxetine and Citalopram on response Inhibition in P Arkinson's disease with clinical and neuroimaging measures. *Hum. Brain Mapp.* **37**(3), 1026–1037. https://doi.org/10.1002/hbm.23087 (2016).
8. Gimenez-Nadal, J. I., Lafuente, M., Molina, J. A. & Velilla, J. Resampling and bootstrap algorithms to assess the relevance of variables: applications to cross section entrepreneurship data. *Empir. Econ.* **56**(1), 233–267. https://doi.org/10.1007/s00181-017-1355-x (2019).
9. Dodge, J., Gururangan, S., Card, D., Schwartz, R. & Smith, N. A. Expected validation performance and estimation of a random variable's maximum. (Accessed 04 Feb 2024) http://arxiv.org/abs/2110.00613 (2021).
10. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci.* **116**(32), 15849–15854,019. https://doi.org/10.1073/pnas.1903070116
11. Kernbach, J. M. & Staartjes, V. E. Foundations of machine learning-based clinical prediction modeling: Part II-generalization and overfitting. In *Machine Learning in Clinical Neuroscience* Acta Neurochirurgica Supplement, vol. 134, (eds Staartjes, V. E. et al.) 15–21. https://doi.org/10.1007/978-3-030-85292-4_3 (Springer International Publishing, 2022).
12. Olaniyi, E. O., Oyedotun, O. K., Ogunlade, C. A. & Khashman, A. In-line grading system for Mango fruits using GLCM feature extraction and soft-computing techniques. *Int. J. Appl. Pattern Recognit.* **6**(1), 58–75. https://doi.org/10.1504/IJAPR.2019.104294 (2019).
13. Benjamin, E. J. et al. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation.* **139**(10), e56–e528 (2019).

14. Arora, S., Santiago, J. A., Bernstein, M. & Potashkin, J. A. Diet and lifestyle impact the development and progression of Alzheimer's dementia. *Front. Nutr.* **10**, https://doi.org/10.3389/fnut.2023.1213223 (2023).
15. Zuhair, M. et al. Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. *Rev. Med. Virol.* **29**(3), e2034. https://doi.org/10.1002/rmv.2034 (2019).
16. Xiong, B., Jiang, W. & Zhang, F. Semi-supervised classification considering space and spectrum constraint for remote sensing imagery. In *2010 18th International Conference on Geoinformatics*, 1–6. https://doi.org/10.1109/GEOINFORMATICS.2010.5567981 (IEEE, 2010).
17. Nadar, N. & Kamatchi, R. A novel student risk identification model using machine learning approach. *Int. J. Adv. Comput. Sci. Appl.* **9**, 305–309. https://doi.org/10.14569/IJACSA.2018.091142 (2018).
18. Khan, A. & Ghosh, S. K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ. Inf. Technol.* **26**(1), 205–240. https://doi.org/10.1007/s10639-020-10230-3 (2021).
19. Yousafzai, B. K., Hayat, M. & Afzal, S. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student, *Educ. Inf. Technol.*, **25**(6), 4677–4697. https://doi.org/10.1007/s10639-020-10189-1 (2020).
20. Smirani, L. K., Yamani, H. A., Menzli, L. J. & Boulahia, J. A. Using ensemble learning algorithms to predict student failure and enabling customized educational paths, *Sci. Program.* **2022**, 1–15. https://doi.org/10.1155/2022/3805235 (2022).
21. Usama, M., Ahmad, B., Xiao, W., Hossain, M. S. & Muhammad, G. Self-attention based recurrent convolutional neural network for disease prediction using healthcare data, comput. *Methods Programs Biomed.* **190**, 105191 (2020).
22. Shukla, N., Hagenbuchner, M., Win, K. T. & Yang, J. Breast cancer data analysis for survivability studies and prediction, *Comput. Methods Programs Biomed.*, **155**, 199–208, https://doi.org/10.1016/j.cmpb.2017.12.011 (2018).
23. Kaur, G. & Chhabra, A. Improved J48 classification algorithm for the prediction of diabetes. *Int. J. Comput. Appl.* **98**(22), 13–17. https://doi.org/10.5120/17314-7433 (2014).
24. Naz, H. et al. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J. Diabetes Metab. Disord.* **19**(1), 391–403 https://doi.org/10.1007/s40200-020-00520-5.
25. Dharma, F. et al. Prediction of Indonesian inflation rate using regression model based on genetic algorithms. *J. Online Inform.* **5**(1), 45–52 https://doi.org/10.15575/join.v5i1.532 (2020).
26. Touzani, S., Granderson, J. & Fernandes, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build.* **158**, 1533–1543 https://doi.org/10.1016/j.enbuild.2017.11.039 (2018).
27. Mohan, S., Thirumalai, C. & Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access.* **7**, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707 (2019).
28. Anuradha, C. & Velmurugan, T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian J. Sci. Technol.* **8**(15). https://doi.org/10.17485/ijst/2015/v8i15/74555 (2015).
29. Hussain, A. A. & Dimililer, K. Student grade prediction using machine learning in iot era. In *International Conference on Forthcoming Networks and Sustainability in the IoT Era*, 65–81. https://doi.org/10.1007/978-3-030-69431-9_6 (Springer, 2021).
30. Mathers, C. D., Boerma, T. & Ma Fat, D. Global and regional causes of death. vol. 92, no. 1, 7–32 (Accessed 14 Jan 2025). https://academic.oup.com/bmb/article-abstract/92/1/7/332071 https://doi.org/10.1093/bmb/ldp028 (Oxford University Press, 2009).
31. Chowdhury, R. et al. Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries. *Eur. J. Epidemiol.* **35**(5), 389–399. https://doi.org/10.1007/s10654-020-00649-w (2020).
32. Townsend, N. et al. Epidemiology of cardiovascular disease in Europe. *Nat. Rev. Cardiol.* **19**(2), 2 https://doi.org/10.1038/s41569-021-00607-3 (2022).
33. Ansari, M. F., Alankar, B. & Kaur, H. A prediction of heart disease using machine learning algorithms. In *Image Processing and Capsule Networks*, vol. 1200, (eds Chen, J. I. Z. et al.) in Advances in Intelligent Systems and Computing, vol. 1200, 497–504. https://doi.org/10.1007/978-3-030-51859-2_45 (Springer International Publishing, 2021).
34. Amarbayasgalan, T., Pham, V. H., Theera-Umpon, N., Piao, Y. & Ryu, K. H. An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets. *IEEE Access.* **9**, 135210–135223. https://doi.org/10.1109/ACCESS.2021.3116974 (2021).
35. Barhoom, A. M., Almasri, A., Abu-Nasser, B. S. & Abu-Naser, S. S. Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms (Accessed 04 Feb 2024) https://philpapers.org/rec/BARPOH-4 (2022).

Acknowledgements

Wish to published.

Author contributions

Authors ContributionYagyanath Rimal: Experimental design, data interpretation, AI Model design-Navneet Sharma: Superior, ML selectionSiddhartha Paudel: Python code Abeer Alsadoon:English proof reading, model validationMadhav Parsad Koirala: Word formattingSumeet Gill: revision.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.