Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER
PUBLISHER

Research Article

Stacking Ensemble Machine Learning Algorithm with an Application to Heart Disease Prediction

Ruhi Fatima[1], **Sabeena Kazi**[2*], **Asifa Tassaddiq**[2], **Nilofer Farhat**[1], **Humera Naaz**[2], **Sumera Jabeen**[3]

[1]Department of Computer Science, College of Computer and Information Sciences, Majmaah University, AlMajmaah, 11952, Saudi Arabia
[2]Department of Basic Sciences and Humanities, College of Computer and Information Sciences, Majmaah University, AlMajmaah, 11952, Saudi Arabia
[3]Department of Computer Science Engineering, CMR Engineering College, Hyderabad, 501401, India
Email: s.badesaheb@mu.edu.sa

Received: 29 January 2023; **Revised:** 16 March 2023; **Accepted:** 20 March 2023

Abstract: Mathematics and statistics have a significant impact on the advancement of most trending sciences like machine learning, artificial intelligence, and data science. In this article, we use the Stacking Ensemble Machine Learning Algorithm (SEMLA) to predict heart disease, considering accuracy (acc), diagnostic odds ratio (Dor), F_1_score, Matthews correlation coefficient (Mcc), receiver operating characteristics-area under curve (roc-auc), and log-loss (log_loss). The data is analyzed using classification learning techniques. We have considered sex, age, cholesterol, fasting blood sugar, the highest rate of heartbeat, type of chest pain, resting electrocardiogram (ECG), angina, depression induced by exercise, peak exercise measurement, major vessel number, a disorder in the blood, and a target attribute to represent the presence and absence of disorders. The approach used allows for the prediction of heart disease and the management of worst-case scenarios. In comparison with the existing models, our proposed model has outperformed other models with an accuracy of 97.28%.

Keywords**:** classifier methods, heart disease, machine learning, prediction, metrics

MSC: MSC62, MSC68

Nomenclature

Acc	Accuracy
Dor	Diagnostic odds ratio
kNN	k-nearest neighbors
LDA	Linear discriminant analysis
LR	Logistics regression
Lr^+	Positive-likelihood ratio
Lr^-	Negative-likelihood ratio
Mcc	Mathews correlation coefficient
MLP	Multi-layer perceptron

DOI: https://doi.org/10.37256/cm.4420232390

Nsvc	Nu-support vector machine
roc-auc	Receiver operating characteristics-area under curve
SEMLA	Stacking Ensemble Machine Learning Algorithm

1. Introduction

Computing technology has made inroads into a variety of industries, including healthcare. As a result, electronic health data could be viewed as a valuable resource in the field of education. Advancements in digital learning techniques are incorporated to better understand possible hidden concerns. Any computational methods used to process healthcare data should be approached with caution. Mathematics is the fundamental foundation of theoretical computer science. Studying and interpreting data pertaining to medical diagnostic processes requires this imperative method [1]. Real experiments are challenging to conduct, as in these instances. These models' and analyses' outputs aid in the understanding of diagnostic techniques and statistically depict various intervention plans and their expected consequences. It is a universally acknowledged fact that important and delicate items handled with care last a long time. Machine learning algorithms can be efficiently utilized in the early diagnosis of diseases to assist doctors in providing the best possible care [2-4]. Machine learning is divided into three categories: supervised, unsupervised, and semi-supervised techniques of learning from data. The goal of being supervised is to achieve the desired result by implementing what has been learned. As the name implies, the second category classifies data without supervision and operates by recognizing patterns in the training data. Finally, the semi-supervised approach combines a few first-type techniques with a large number of second-type techniques to take advantage of both.

The heart is a vital organ in the anatomy of living creatures, and it is important for maintaining a healthy lifestyle. As per data compiled from 1999 to 2020, heart disease is rated as the number one cause of death [5]. According to WHO figures, roughly 17.9 million individuals died in the year 2019 as a result of heart-related problems [6]. A closed fist-sized organ is the first functioning portion of the remnant in the embryo. Blood must be pumped from the most amazing organ in the body to all other key systems in order for the entire body to work as it should. The high fatality rate, even in advanced nations, is explained by its malfunction. For the human body to function properly, the heart must be in good health. Smoking, obesity, high blood glucose, and a lack of physical activity are a few major risk factors for heart disease. Unlike in ancient times, when there was a scarcity of anatomical and pathological information on issues relating to the heart, there are now a plethora of tests and tools available to assess the organ's health. Specialists diagnose cardiac illness based on a patient's medical history and the results of tests such as 2Decho and ultrasound. The disease could be caused by a variety of factors, including fat in the arteries, an irregular heartbeat, a flaw in the organ itself, its muscles, or infections, as well as abnormalities in the heart regions. Symptoms will vary depending on the problem. One illness's symptoms may or may not be the same as another's. A person with chest pain, breathing problems, or fainting should hurry to the hospital since prompt medical attention could save their lives. Cardiac biomarkers are also used to assess the health of the heart. When myocardial necrosis occurs, such as in myocardial infractions, cardiac enzymes are released into the circulation. The most notable cardiac biomarkers include myoglobin, troponin, and creatinine kinase. Troponin I and Troponin II are released into the bloodstream between 3 and 4 hours after a myocardial infarction, which can be detected approximately 10 days before the damage to the heart occurs.

This study aims to utilize learning approaches and propose a model based on patient data to allow the prediction of heart disease and the management of worst-case scenarios. The proposed model receives the set of attributes as input and processes it to automate the target variable, giving better results in comparison to the other learning methods.

2. Materials and methods

2.1 *Dataset and preprocessing*

The public data on heart disease [7] is utilized to apply the models, confirming that all methods were performed according to relevant guidelines or regulations and applying machine learning models, which are briefly discussed. The data comprises 1,025 records, which are preprocessed to be complete without any missing data. By choosing a subset of a population at random, simple random sampling is used to draw statistical conclusions about the population. Each person in the population has an exactly equal probability of being chosen using this sampling technique. Using the fact that the data considered in this paper is complete and a sample size calculator with a 99% confidence level, the ideal size of the sample was 822 records. So, we proceed with a simple random sample of 820 of 1,025 records comprising 80% of the data with 14 features: chest pain type, age, sex, fasting blood sugar (FBS), serum cholesterol (S.cholesterol), resting blood pressure (resting-BP), exercise-induced angina, achieved maximum heart rate (max. HR), resting electrocardiogram (resting-ECG), number of major vessels, thalassemia, the slope of peak exercise segment (slope), segment-depression, and target. The target distribution contains 47.9% with no disease and 52.1% with a disease, indicating that the data is balanced. The entries for each attribute are listed below. In the chest pain type, 1 indicates conventional angina, 2 atypical angina, 3 non-anginal pain, and 4 non-symptomatic pain. The feature age is the number of years in the attribute sex, where male is represented by 1 and female is represented by 0. The slope values are 1 (upsloping), 2 (flat), and 3 (downsloping). The values for S.cholesterol and resting BP are in mg/dl and mmHg, respectively. When the FBS value exceeds 120 mg/dl, it is 1; otherwise, it is 0. The resting-ECG value is 0 for normal, 1 for abnormal, and 2 for definite left ventricular hypertrophy criteria. The exercise-induced angina value is 1 for yes and 0 for no, while the number of major vessels displays 0-3 colored by fluoroscopy. Finally, thalassemia values denote defects with 1 fixed fault, 2 normal defects, and 3 reversible flaws. The variable target is set to 0 if the patient has no disease and 1 otherwise. Figure 1 shows a block diagram of the methodology used in this study.

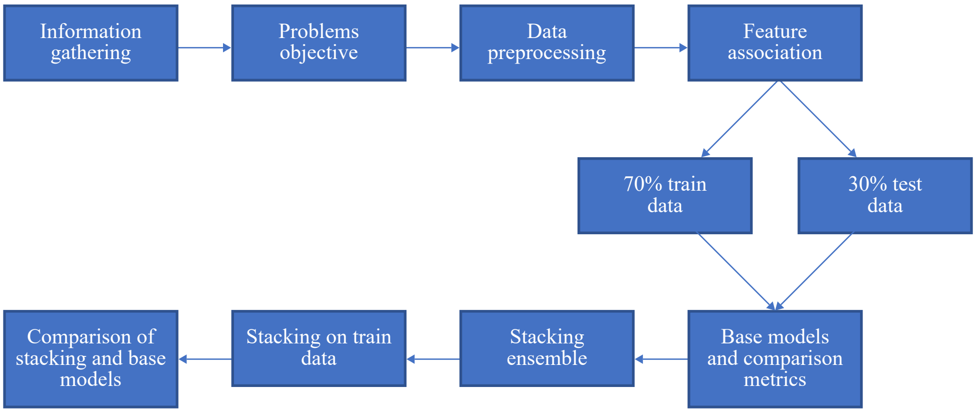


Figure 1. Block diagram

Table 1 shows the descriptive statistics for the dataset's integer and float-type data observations. Each numeric feature's mean, maximum, minimum, standard deviation, and standard error were included.

Table 1. Descriptive statistics of numeric features of data

	Mean	Std. deviation	Max.	Min.	Std. error
Age	54.45	9.2	77	29	0.32
Resting-BP	131.72	17.88	200	94	0.62
S.cholesterol	247.36	52.65	564	126	1.84
Max. HR achieved	149.13	23.02	202	71	0.8
Segment-depression	1.06	1.18	6.2	0	0.04
No. of major vessels	0.76	1.04	4	0	0.04

The statistics of categorical features (binary and multi-category) are shown in Table 2.

Table 2. Descriptive statistics of categorical features of data

	Mean	Std. deviation	Max.	Min.	Std. error
Sex	0.68	0.47	1	0	0.016
Chest pain type	0.97	1.02	3	0	0.036
FBS	0.16	0.36	1	0	0.013
Resting-ECG	0.51	0.52	2	0	0.018
Exercise-induced angina	0.34	0.47	1	0	0.017
Slope (peak exercise segment)	1.39	0.62	2	0	0.022
Thalassemia	2.3	0.62	3	0	0.022

The 14 data attributes are classified as numerical, binary, and category features during preprocessing. Age, S.cholesterol, resting-BP, max. HR achieved, segment-depression, and the number of major vessels are all numerically grouped attributes. The binary features are sex, FBS, exercise-induced angina, and target, whereas the categorical features are the rest. With only two categories, the binary features are likewise essentially categorical. The features are classified using the data binning technique [8]. To determine the relationship between the properties of the chosen data, the correlation is explored. Point-biserial correlation is used for numerical features, whereas Cramer's V [9] is for categorical data. Figure 2 shows that numerical features have a weak association with the target.

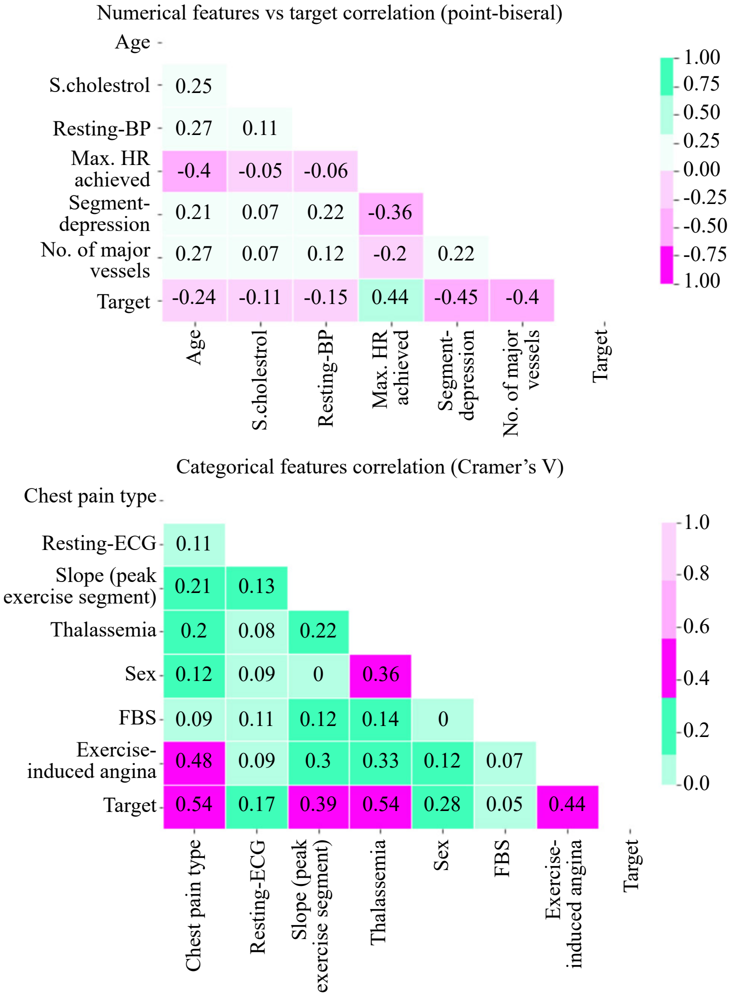


Figure 2. Correlation

We utilized a Jupyter notebook to implement the chosen methods in the Python programming language. The Scikit-Learn library was used to get the result of the metric under evaluation.

2.2 *Classification methods*

Classification is an important branch of machine learning to categorize data and use it effectively and efficiently. On the classified data, complex and varied actions in many different fields can be performed to get better insights for further analysis [10]. We chose the strategies from a vast variety of classification algorithms [11, 12]. They are LR [13], kNN [14], Nsvc [15], LDA [16], and MLP [17].

2.2.1 *Proposed SEMLA*

A meta-model linearly pools the results from many base models to be able to advance the functioning of machine learning [18]. A top blending layer is extended to the voting ensemble technique to acquire the superlative aggregation of the models in consideration. Ensemble-learning predictive models improve poor classifiers' efficiency, statistics, and computation performance. Stacking can be utilized to increase performance in approaches including optimal feature selection, incremental and nonstationary learning, mistake detection, error correction, and decision confidence improvement [19]. The basic process of stacking is to train and make predictions on the original training dataset using first-level learners. These predictions are then combined and make up the training data for the meta-learner. That is, the meta-learner uses the output of the first-level learners as input. Although it is feasible to generate stacked ensembles out of the same learning algorithms, first-level learners are frequently made up of various and diverse learning algorithms. As depicted in Figure 3, the selected five base models are implemented on the partitioned heart data. The predicted values are represented as P1 of LR, P2 of kNN, P3 of LDA, P4 of Nsvc, and P5 of MLP. The five predictions are aggregated in SEMLA to produce the final predicted value (P).

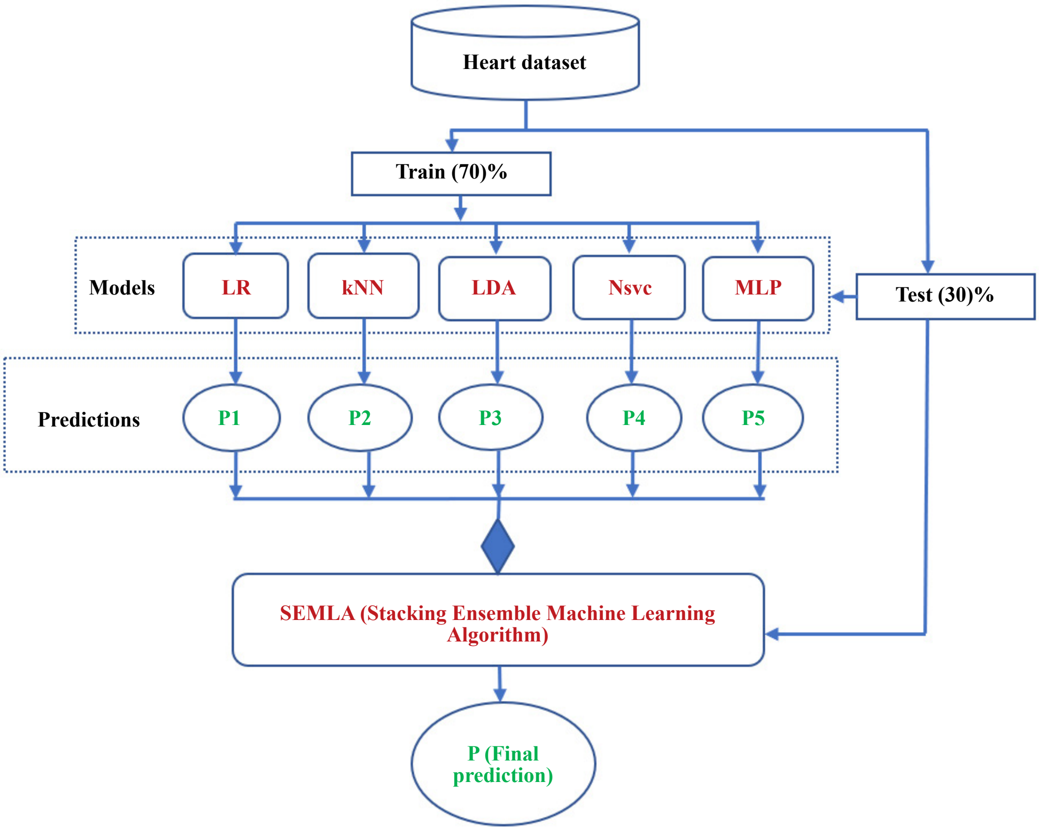


Figure 3. SEMLA-model

2.2.2 *Pseudocode*

```
Input: Training data
Output: Final prediction
Start
Step 1: Read the dataset
Step 2: Univariate feature selection (point-biserial, Cramer’s V)
Step 3: Partition the data
Step 4: Build level-0 models
for each model in base models
    Append it to the level-0 classifier
end for
Step 5: Define level-1 classifier
Step 6: Build the stacking classifier using level-0, and level-1 estimators, and cross-validation
Step 7: Fit the model on train data
Step 8: Predict the model on train and test data
Step 9: Go to Step 3 (for categorical data)
Step 10: Else stop
```

2.2.3 *Basic code for SEMLA*

```
level0 = list()
level0.append((‘LR’, LogisticRegression()))
level0.append((‘kNN’,KNeighborsClassifier()))
level0.append((‘LDA’, LinearDiscriminantAnalysis()))
level0.append((‘Nsvc’, NuSVC()))
level0.append((‘MLP’, MLPClassifier()))
level1 = RandomForestClassifier() # meta learning model
model = StackingClassifier(estimators = level0, final_estimator = level1, cv = 5)
predict_st = model.predict(X_train)
accuracy_st = metrics.accuracy_score(y_train, predict_st)
print(“Accuracy :train- Stacking = ”, accuracy_st)
pred_prob_modt = model.predict_proba(X_train)
predict_s = model.predict(X_test)
accuracy_s = metrics.accuracy_score(y_test,predict_s)
print(“Accuracy :test- Stacking = ”, accuracy_s)
pred_prob_mod = model.predict_proba(X_test)
```

3. Results

3.1 *Metrics for model evaluation*

The performance of the model is evaluated using metrics. It is very important to select the appropriate metric [20, 21]. The following measures have been used to evaluate the efficiency of the base model and the proposed model.

3.1.1 *Confusion matrix*

It is also known as a contingency table and is used to summarize the chosen five classifiers’ performance [22]. Table 3 shows the tabulated confusion matrix for the train and the test set following data preprocessing. The following notations are used:

T_P (True_Positive): heart disease is anticipated in patients who have it.

T_N (True_Negative): patients without heart disease are anticipated to be free of disease.

F_P (False_Positive): the patients without heart disease are anticipated to develop it.

F_N (False_Negative): heart disease patients are anticipated to be free of the condition.

It is evident from the values in Table 3 that the classification summary of SEMLA is better in comparison to other models on both train and test data. SEMLA has correctly classified training data as 278 patients having heart disease and 294 as patients not having the disease. The summarized classification result on test data for SEMLA, with 109 true-positives and 130 true-negatives, is better than other models. The performance of the MLP is also quite good in comparison with LR, LDA, and kNN. Model Nsvc is a poor classifier among all the models on training and testing data.

Table 3. Confusion matrix of the train, test data

	Train				Test				Total records
Model	T_P	F_P	F_N	T_N	T_P	F_P	F_N	T_N	
LR	241	38	24	271	97	17	6	126	820
kNN	241	38	34	261	100	14	9	123	820
LDA	240	39	26	269	97	17	6	126	820
Nsvc	250	29	34	261	101	13	17	115	820
MLP	255	24	13	282	102	12	3	129	820
SEMLA	278	1	1	294	109	5	2	130	820

3.1.2 *Sensitivity*

A metric [23] uses the model to describe the number of real positive cases that were projected to be positive. In other words, it is regarded as a true-positive rate, which measures the likelihood that a condition associated with heart disease would be accurately detected using equation 1.

$$sensitivity = \frac{T_p}{T_p + F_N} \tag{1}$$

3.1.3 *Specificity*

It is described as the model's ability to accurately forecast T_N instances. It is the inverse of recall and is referred to as T_N rate. It counts the no heart disease individuals who were correctly recognized as healthy people without the disease with equation 2.

$$specificity = \frac{T_N}{T_N + F_P} \tag{2}$$

3.1.4 *Accuracy*

A metric is used to show the relationship between the measured result and the actual value. It is used to quantify the ability to differentiate between a person with the disease and someone who does not have the disease. Accuracy can be determined using sensitivity and specificity, with prevalence results [24] using equation 3.

$$accuracy = \frac{T_p + T_N}{T_p + T_N + F_p + F_N} \tag{3}$$

The above-mentioned metrics are used to measure the performance of the models, and the results achieved are recorded in Tables 4, and 5. Table 4 demonstrates the obtained values of metric sensitivity, specificity, and accuracy using 13 features on train (sens_1_train, spec_1_train, acc_1_train) and test (sens_1_test, spec_1_test, acc_1_test) data. Table 5 records the results of metrics with categorical features on the train (sens_2_train, spec_2_train, acc_2_train) and test (sens_2_test, spec_2_test, acc_2_test) data. SEMLA has obtained sensitivity, specificity, and accuracy of 98.4%, 96.21%, and 97.28%, respectively, which is better than other models. And also, it is evident from Table 5 that SEMLA implemented with only categorical features has outperformed other models by achieving 86.4%, 90.91%, and 88.72% sensitivity, specificity, and accuracy, respectively.

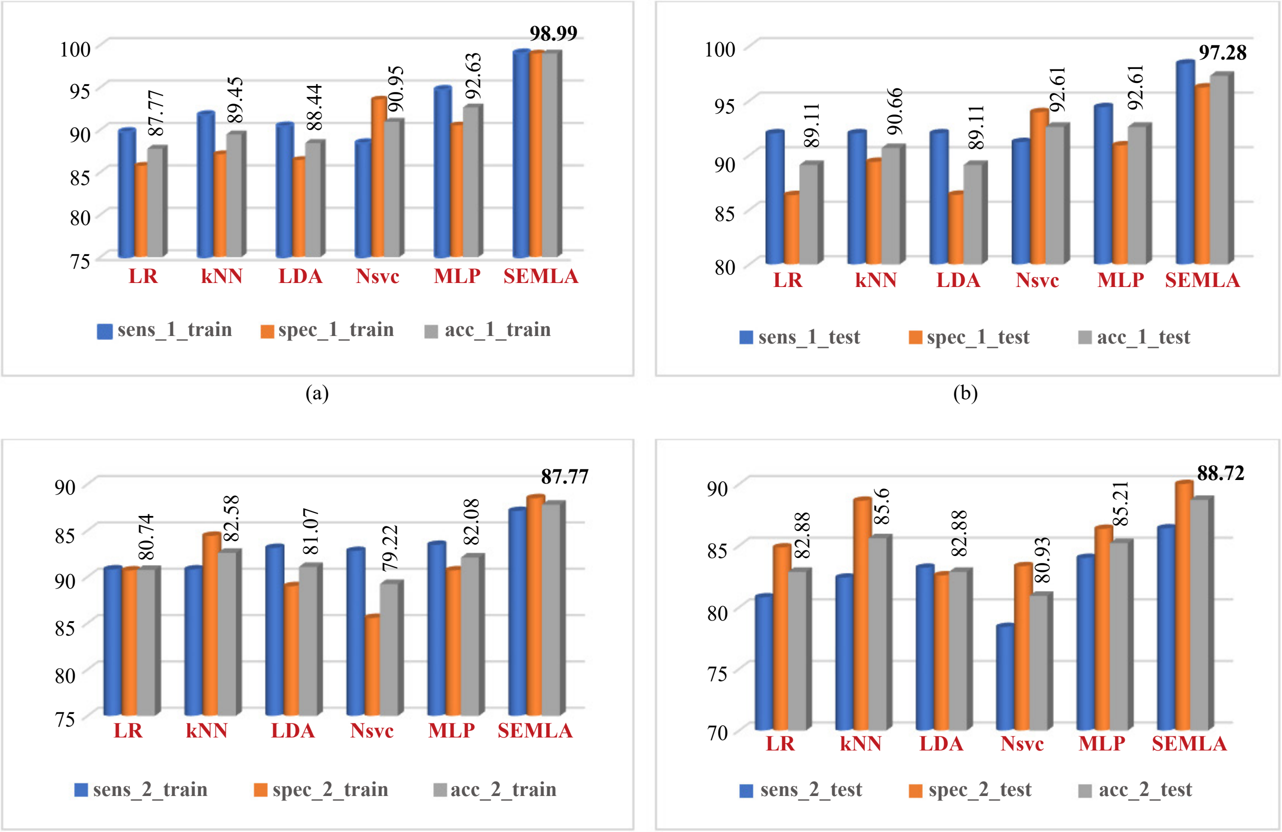
Table 4. Sensitivity, specificity, the accuracy of the train, and test data with all features

		Train			Test	
Model	sens_1_train	spec_1_train	acc_1_train	sens_1_test	spec_1_test	acc_1_test
LR	89.73	85.76	87.77	92	86.34	89.11
kNN	91.72	87.12	89.45	92	89.39	90.66
LDA	90.4	86.44	88.44	92	86.37	89.11
Nsvc	88.41	93.56	90.95	91.2	93.94	92.61
MLP	94.7	90.51	92.63	94.4	90.9	92.61
SEMLA	**99.0**	**98.98**	**98.99**	**98.4**	**96.21**	**97.28**

Table 5. Sensitivity, specificity the accuracy of the train, and test data with the categorical features

		Train			Test	
Model	sens_2_train	spec_2_train	acc_2_train	sens_2_test	spec_2_test	acc_2_test
LR	80.79	80.68	80.74	80.8	84.85	82.88
kNN	80.79	84.41	82.58	82.4	88.64	85.60
LDA	83.11	78.98	81.07	83.2	82.58	82.88
Nsvc	82.78	75.56	79.22	78.4	83.33	80.93
MLP	83.44	80.69	82.08	84	86.36	85.21
SEMLA	**87.09**	**88.47**	**87.77**	**86.4**	**90.91**	**88.72**

Tables 4 and 5 are visualized in Figures 4 (a), (b), (c), and (d). It is aesthetically evident that the accuracy achieved by SEMLA is better than other models, both for training and testing data. Figures 4 (a) and (b) depict the results of models on the train and test data with 13 features. Figures 4 (c) and (d) show the metric values obtained with categorical train and test data. SEMLA has performed incredibly well when it comes to train data. Furthermore, compared to categorical features, the values of data containing all features are better. With all features, the train's accuracy is greater than the accuracy of the test data, indicating overfitting of the model; however, with categorical features, it is optimal.



(a)

(b)

(c)

(d)

Figure 4. Sensitivity, specificity, accuracy: (a) train data with 13 features and target; (b) test data with 13 features and target; (c) train data with categorical features and target; (d) test data with categorical features and target

3.1.5 Lr^+

A metric in [25] was used to quantify the probability of patients who have heart disease being diagnosed positively over the probability of patients who are negatively diagnosed with the disease. In equation 4, '+' represents the increase in the odds of the disease when the diagnostic test is positive.

$$Lr^+ = \frac{sensitivity}{1 - specificity} \tag{4}$$

3.1.6 Lr^-

This measurement quantifies the ratio of the probability of patients being diagnosed with heart disease negatively to the probability of people not having the disease. '–' indicates the decrease in the odds of the disease when the diagnostic test is negative, in equation 5.

$$Lr^- = \frac{1 - sensitivity}{specificity} \tag{5}$$

3.1.7 Dor

The effectiveness of the diagnosis of disease data can be measured using the metric Dor [26]. It is a summary of likelihood (Lr^+, Lr^-) ratios. It is calculated using "equation 6". The value of the metric ranges from zero to infinity.

The result of the metric is interpreted as less than zero, specifying that the outcome should be inverted; exactly one signifies a positive prediction irrespective of the actual value; and greater than one signifies a higher performance of the prediction.

$$Dor = \frac{sensitivity * specificity}{(1 - sensitivity)(1 - specificity)} \tag{6}$$

The results obtained for the above metrics are shown in Tables 6 and 7. Table 6 demonstrates the output of metrics Lr^+ (Train_Lr^+, Test_Lr^+), Lr^- (Train_Lr^-, Test_Lr^-), (Train_Dor_1, Test_Dor_1) applied to 13 features. Table 7 shows the output achieved by the metrics (Train_Lr^+,Test_Lr^+,Train_Lr^-, Test_Lr^-, Train_Dor_2, Test_Dor_2) on categorical features of data.

Table 6. Lr^+, Lr^-, Dor_1 of train, test data with 13 features

		Train			Test	
Model	Train_ Lr^+	Train_ Lr^-	Train_Dor_1	Test_ Lr^+	Test_ Lr^-	Test_Dor_1
LR	6.303	0.119	52.66	6.747	0.093	72.83
kNN	7.121	0.095	74.94	8.674	0.089	96.93
LDA	6.667	0.111	60.01	6.747	0.093	72.83
Nsvc	13.727	0.124	110.82	15.048	0.094	160.64
MLP	9.978	0.059	170.45	10.384	0.062	168.57
SEMLA	**97.357**	**0.01**	**9700.89**	**25.978**	**0.017**	**1562.1**

Table 7. Lr^+, Lr^-, Dor_2 of train, test data with categorical features

		Train			Test	
Model	Train_ Lr^+	Train_ Lr^-	Train_Dor_2	Test_ Lr^+	Test_ Lr^-	Test_Dor_2
LR	4.181	0.238	44.39	5.323	0.226	23.57
kNN	5.181	0.228	38.83	7.251	0.199	36.52
LDA	3.955	0.214	38.59	4.775	0.203	23.47
Nsvc	3.392	0.228	26.37	4.704	0.259	18.15
MLP	4.319	0.205	89.57	6.160	0.185	33.25
SEMLA	**7.556**	**0.146**	**2221.03**	**9.504**	**0.149**	**63.53**

The Dor readings of Tables 6 and 7 are visualized to highlight the performance of SEMLA in comparison to the selected five models in Figure 5.

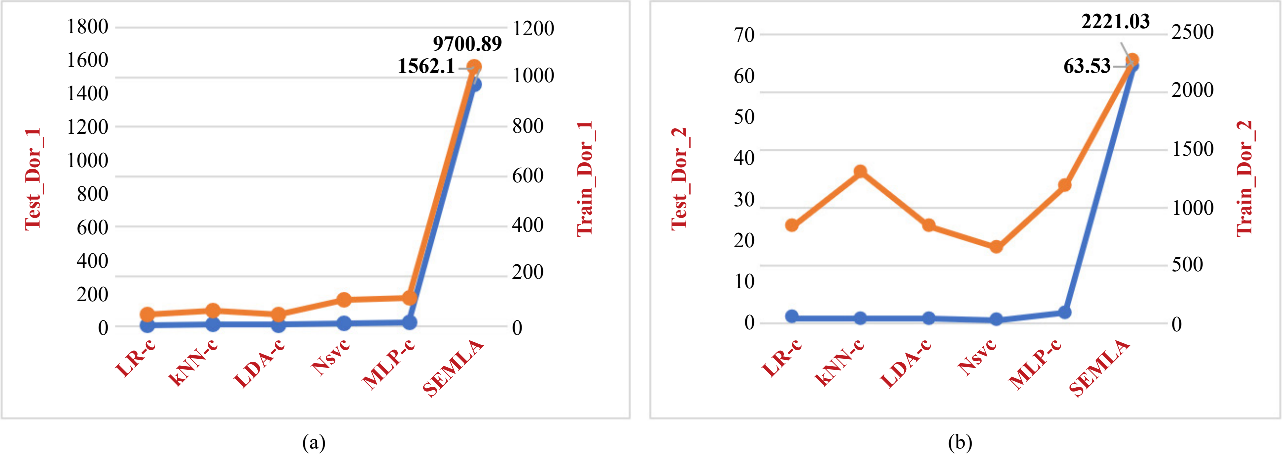


Figure 5. Dor Evaluation: (a) train, test data with 13 features; (b) train, test data with categorical features

3.1.8 *Precision and recall*

This metric [27] is used to qualitatively measure different models' performance. As in equation 7, it is calculated as the ratio of true-positive and (sum of true-positive and false-positive).

$$precision = \frac{T_p}{T_p + F_p} \tag{7}$$

Recall gives a quantitative measure of machine learning models. Equation 8 is used to find the recall by dividing the true positive by (the true positive and false negative).

$$recall = \frac{T_p}{T_p + F_N} \tag{8}$$

3.1.9 *F_1_score*

A weighted average of recall and precision can be calculated using this metric. F_1_score can be computed using equation 9. It can also be defined as the harmonic mean of sensitivity and precision. Its best value is one, whereas its worst value is zero.

$$F_1\_score = \frac{2T_p}{2T_p + F_p + F_N} \tag{9}$$

3.1.10 *Mcc*

The Pearson product-moment correlation coefficient between actual and predicted values is calculated using a contingency matrix method known as the Mcc and can be computed with equation 10. A coefficient of +1.0 denotes a perfect prediction; 0.0 denotes a prediction that is no better than chance; and -1.0 denotes the worst possible prediction.

$$Mcc = \frac{T_p T_N - F_p F_N}{\sqrt{(T_p + F_p)(T_p + F_N)(T_N + F_p)(T_N + F_N)}} \tag{10}$$

In Table 8, pre_1_tr, re_1_tr, f1score_1_tr, and Mcc_1_tr represent the precision, recall, F_1_score, and Mcc of the models on 70% of the training data, and pre_1_t, re_1_t, f1score_1_t, and Mcc_1_t are used for 30% of the test data with 13 features. Table 9 shows the values of the above metrics (pre_2_tr, re_2_tr, f1score_2_tr, and Mcc_2_tr) for train data and (pre_2_t, re_2_t, f1score_2_t, and Mcc_2_t) for test data with categorical attributes.

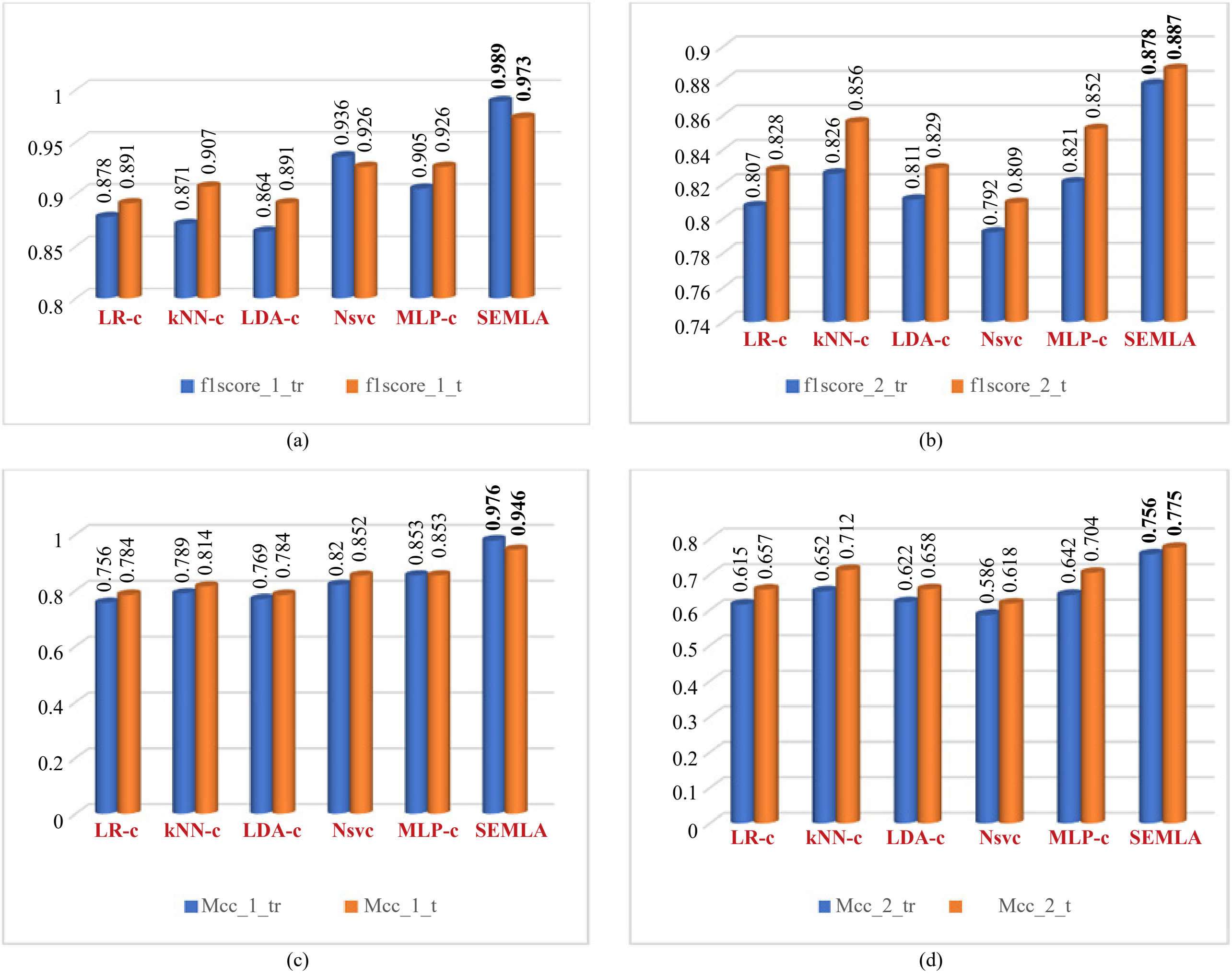
Table 8. Precision, recall, F_1_score of the train, test data with 13 features

	Train				Test			
Model	pre_1_tr	re_1_tr	f1score_1_tr	Mcc_1_tr	pre_1_t	re_1_t	f1score_1_t	Mcc_1_t
LR	0.866	0.897	0.878	0.756	0.865	0.920	0.891	0.784
kNN	0.879	0.917	0.871	0.789	0.891	0.920	0.907	0.814
LDA	0.872	0.904	0.864	0.769	0.865	0.920	0.891	0.784
Nsvc	0.934	0.884	0.936	0.820	0.934	0.912	0.926	0.852
MLP	0.911	0.947	0.905	0.853	0.908	0.944	0.926	0.853
SEMLA	**0.990**	**0.990**	**0.989**	**0.979**	**0.961**	**0.984**	**0.973**	**0.946**

Table 9. Precision, recall, F_1_score of the train, and test data with categorical features

	Train				Test			
Model	pre_2_tr	re_2_tr	f1score_2_tr	Mcc_2_tr	pre_2_t	re_2_t	f1score_2_t	Mcc_2_t
LR	0.811	0.808	0.807	0.615	0.835	0.808	0.828	0.657
kNN	0.841	0.808	0.826	0.652	0.873	0.824	0.856	0.712
LDA	0.802	0.831	0.811	0.622	0.819	0.832	0.829	0.658
Nsvc	0.776	0.828	0.792	0.586	0.817	0.784	0.809	0.618
MLP	0.816	0.834	0.821	0.642	0.854	0.840	0.852	0.704
SEMLA	**0.886**	**0.871**	**0.878**	**0.756**	**0.90**	**0.864**	**0.887**	**0.775**

Figure 6 represents the visualized entries of Tables 8 and 9 (F_1_score and Mcc) separately. The SEMLA scored 0.989 on train data and 0.973 on test data using 13 features, which is an excellent F_1_score. Similarly, the SEMLA did well with categorical features on the train and test data, scoring 0.878 and 0.887, respectively. The model also scored higher on Mcc for train and test with 13 features (0.979, 0.946) and for train and test with categorical (0.756, 0.775).



(a)

(b)

(c)

(d)

Figure 6. F_1_score, Mcc evaluation: (a) F_1_score of the train, test data with 13 features; (b) F_1_score of the train, test data with categorical features; (c) Mcc of the train, test data with 13 features; (d) Mcc of the train, test data with categorical features

3.1.11 *roc_auc score*

The receiver operating characteristics (roc) metric is used to evaluate binary classification types, and the area under the curve (auc) summarizes the area under the curve. The positive and negative class points are perfectly distinguished [28] when the auc value is 1. When auc is 0, all negatives are predicted as positives and all positives as negatives. The rauc_score_1 with 13 features and the rauc_score_2 considering categorical features are recorded in Table 10. SEMLA has overtaken with 0.998 on train data and 0.992 on test data with 13 features. And also, SEMLA has achieved 0.933 on train data and 0.929 on test data with the categorical features.

Table 10. roc_auc score of the train, test data

Model	Train, test data with 13 features		Train, test data with categorical features	
	rauc_score_1_train	rauc_score_1_test	rauc_score_2_train	rauc_score_2_test
LR	0.951	0.953	0.889	0.889
kNN	0.944	0.946	0.918	0.927
LDA	0.951	0.951	0.889	0.889
Nsvc	0.977	0.984	0.5	0.5
MLP	0.973	0.967	0.903	0.903
SEMLA	**0.998**	**0.992**	**0.933**	**0.929**

3.1.12 *auc*

In 1971, [29] the roc curve was introduced in medicine for a radiologist to apply different decision criteria based on the contrast of the percentage of true-positive against false-positive diagnoses. The discriminative capacity of predictive models is commonly assessed using the auc. The risk distribution of patients with the disease and people with NO disease can be presented using the roc plot in Figure 7. The difference in risk between patients with and without disease is shown by the area between the roc curve and diagonal. The higher the auc, the greater the area between roc and diagonal, representing a higher separation between the risk of disease and NO disease people. The area under the curve is obtained by plotting (1-sensitivity) on the *x*-axis and sensitivity on the *y*-axis.

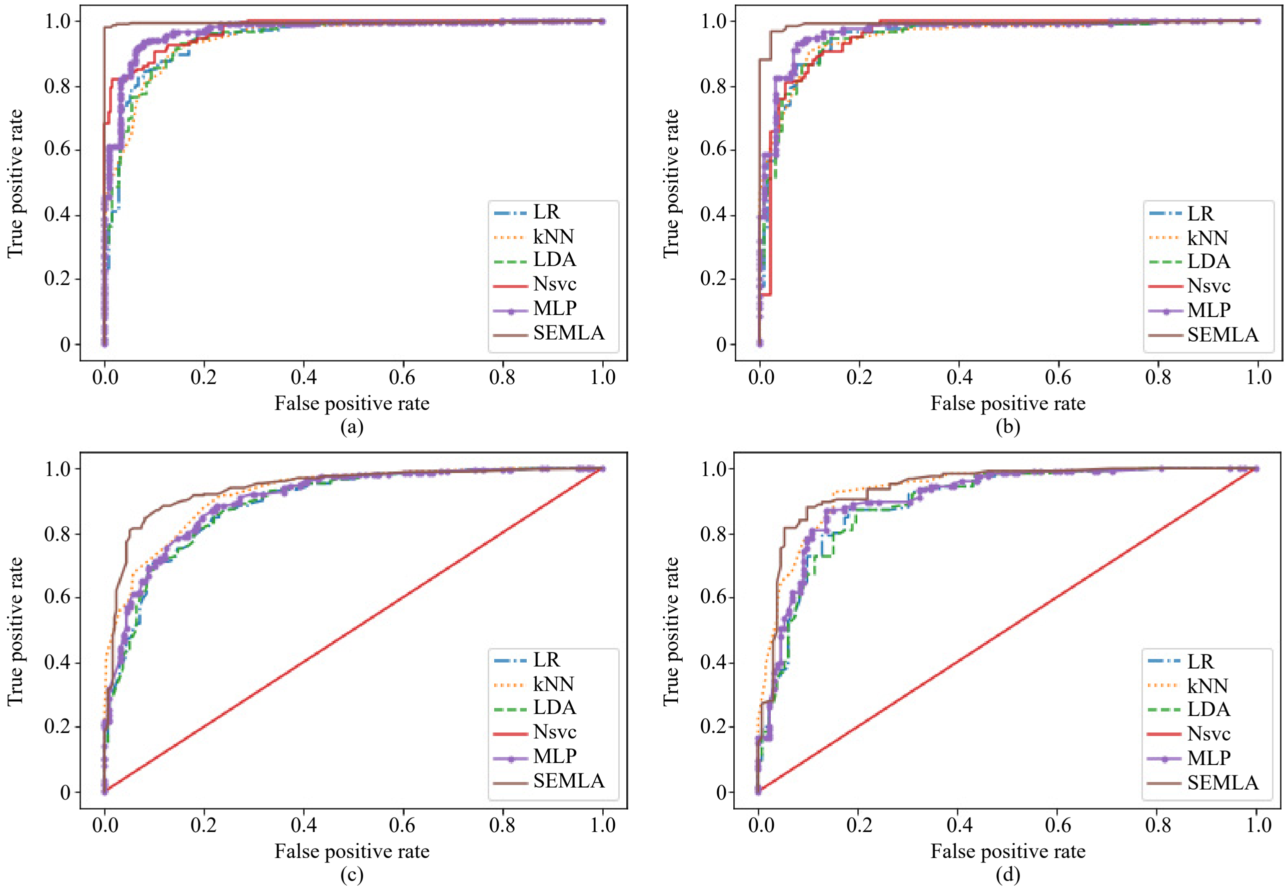


Figure 7. roc_auc_curve: (a) train data with 13 features; (b) test data with 13 features; (c) train data with categorical features; (d) test data with categorical features and target

3.1.13 *Log_loss*

There are many standard loss functions like Brier-loss, spherical-loss, and logarithmic-loss that can be used for probabilistic predictive problems. The log_loss is the most used metric [30, 31]. It is a probability-based metric used for the comparison of the classifier to measure the prediction using equation (11). The values of the metric range from 0 to 1, inclusive. Lower log_loss represents efficient model prediction, value 0 is obtained for a perfect model

$$\text{log\_ loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right)\log\left(1 - p\left(y_i\right)\right)\right) \tag{11}$$

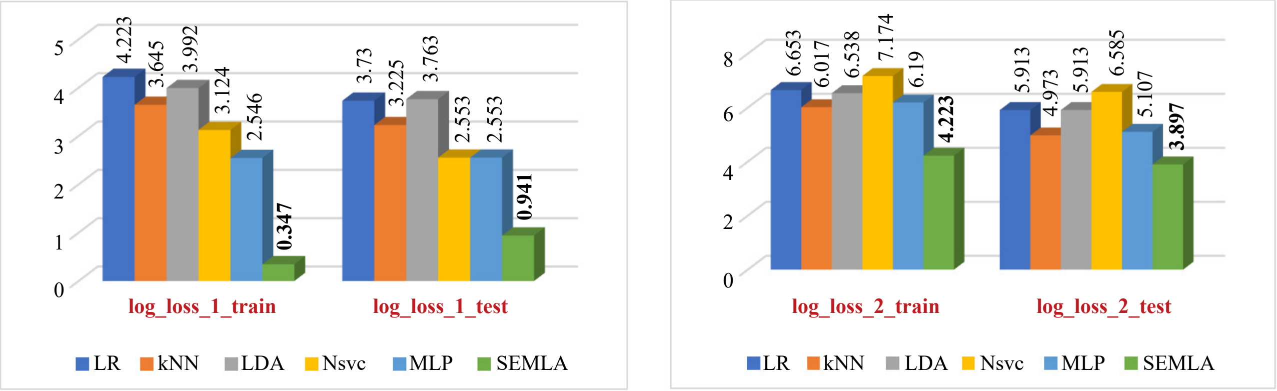
where p is probability prediction.

In Table 11, the values obtained for the log_loss metric are recorded. The log_loss_1 (train, test) values result from all 13 features of the data, and log_loss_2 (train, test) records the reading of metrics with categorical features of the data.

Table 11. Log_loss of the train, test data with 13 features, and categorical features

	Train, test data with 13 features		Train, test data with categorical features	
Model	log_loss_1_train	log_loss_1_test	log_loss _2_train	log_loss _2_test
LR	4.223	3.73	6.653	5.913
kNN	3.645	3.225	6.017	4.973
LDA	3.992	3.763	6.538	5.913
Nsvc	3.124	2.553	7.174	6.585
MLP	2.546	2.553	6.190	5.107
SEMLA	**0.347**	**0.941**	**4.223**	**3.897**

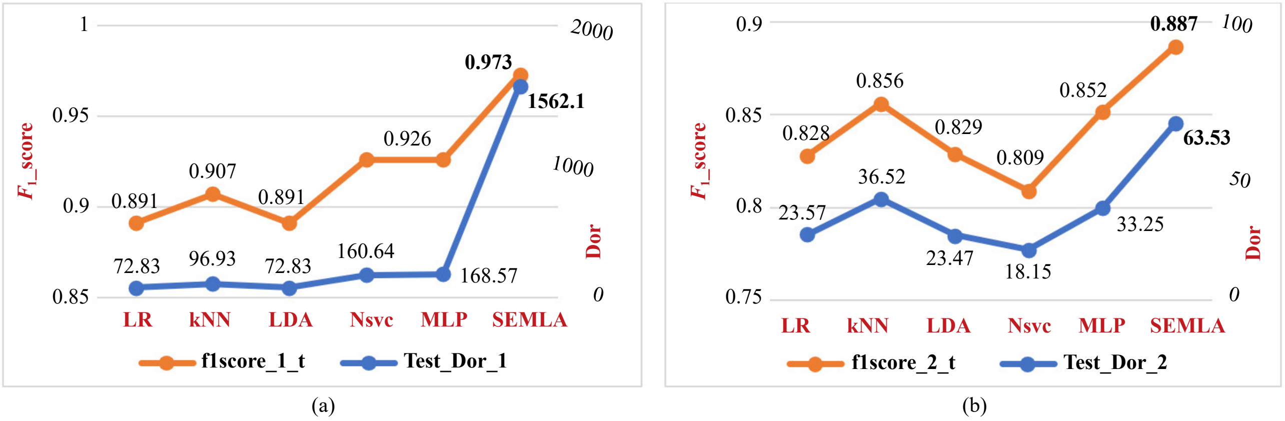
The results of the log_loss function on train and test data with 13 features obtained in Table 11 are visualized in Figure 8 (a), and results with categorical features are shown in Figure 8 (b). The loss in classification using SEMLA is very small in comparison to other models. Only 0.347% and 0.941% losses occurred on the train, test data with 13 features. Similar results were observed with categorical feature data.



(a) (b)

Figure 8. log_loss evaluation: (a) train, test data with 13 features; (b) train, test data with categorical features

As further analysis [32, 33], the findings of the F_1_score and Dor with all features are visualized in Figure 9 (a), and in Figure 9 (b), the results obtained for the F_1_score and Dor with categorical features are shown.

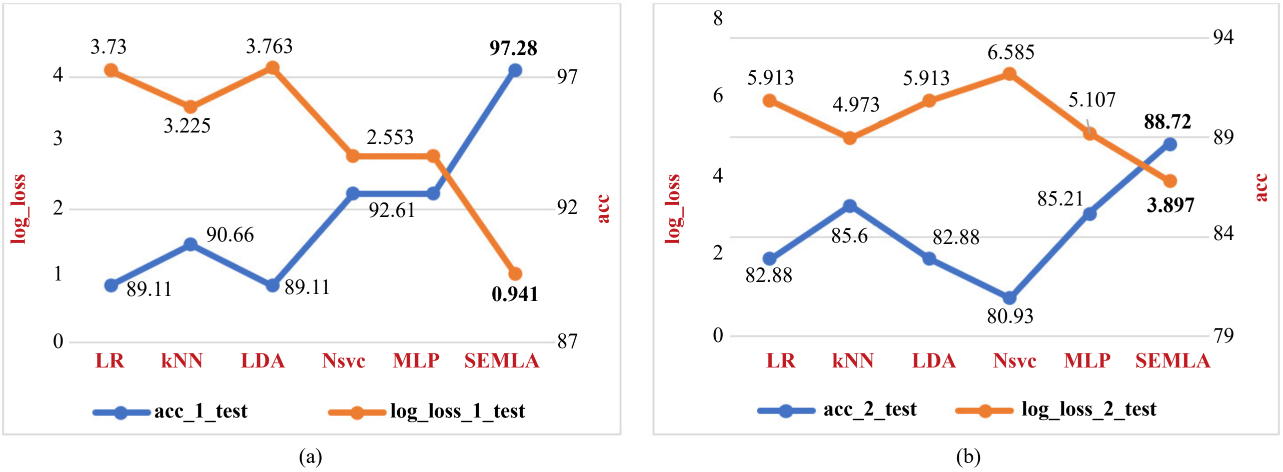


(a)

(b)

Figure 9. Dor, F_1_score combo evaluation: (a) test data with 13 features; (b) test data with categorical features

The findings of accuracy and log_loss with 13 features and with categorical features on test data are summarized in Figures 10 (a) and (b). The SEMLA has achieved excellent accuracy, resulting in a minimum loss.



(a)

(b)

Figure 10. Accuracy, log_loss combo evaluation: (a) test data with 13 features; (b) test data with categorical features

4. Discussion

Several researchers obtained a survey [34] of research publications on the prediction of heart disease and accuracy, f-measure. To disclose the information in the dataset, quantitative data analysis is used [35, 36]. The authors of [37] looked at major factors that contribute to cardiovascular diseases, such as dyslipidemia, hypertension, diabetes, smoking, and lack of physical activity. To improve the system's efficiency, a web application was created [38, 39] to collect patient data and implement several machine learning techniques. To increase accuracy, "Hybrid Random Forest with a Linear Model (HRFLM)" [40] combines the advantages of random forest (RF) and linear-method. Support vector machine (svm), kNN, naïve-Bayes (NB), decision-tree (dtree), and adaboost were demonstrated to quantify regressor metrics using various models in [41]. On classification methods such as neural networks (NN), dtree, svm, LR, LDA, RF, kNN, and NB [42], holdout, stratified k-fold, k-fold cross-validation, and repeated random procedures were used, and an accuracy of 71.82% was achieved using NN with holdout cross-validation. The voting model [43] was implemented with 84.1% accuracy using svm and dtree. [44] used Weka to illustrate classification algorithms (svm, NB, dtree, kNN) that obtained 84.33% accuracy. In [45], a statistical correlation was employed to choose the features to increase accuracy by utilizing 18 distinct classifiers and achieving an efficiency of 85%. [46] recorded the performance of kNN for k values ranging from 1 to 20 and found an accuracy of 87%, while [47] found an efficiency of 90.79% with a k value of 7. In contrast, RF in [48] has a superior accuracy of 91.80% when compared to other models. In [49], a cost-sensitive ensemble model was demonstrated to attain 92% accuracy using Matlab employing statistical t-test comparison. Fast conditional mutual information (fcmim) feature selection was compared to various feature selection methods in [50], and it attained a percentage of 92.37 accuracies in 0.001 seconds. [51] illustrates four distinct preprocessing strategies for dealing with missing values. To achieve 95.83% accuracy, boosting techniques such as extreme and adaptive gradient, light gradient boosting, extra trees, stochastic gradient descent, nsvc, and stacking were used. To demonstrate the findings with and without principal component analysis, an artificial neural network model with embedded regularization based on standard deviation [52] was built, with an accuracy of 96.3%. An accuracy of 96.3% was achieved [53] by using a voting classifier with svm, RF, LR, XGBoost, and deep learning (Convolution-NN, Deep-NN) as base models. The machine learning model was implemented after applying Pearson correlation, recursive elimination, analysis of variance (ANOVA), least absolute shrinkage and selection operator (LASSO), and decision tree (DT) feature selection techniques. The authors [54] have preprocessed the data and applied LASSO for feature selection, then sequential deep learning with a fully connected dense layer to achieve an accuracy of 94.2%.

The accuracy of SEMLA compared with other models studied in the literature is shown in Figure 11. SEMLA has outperformed other models with an accuracy of 97.28%.

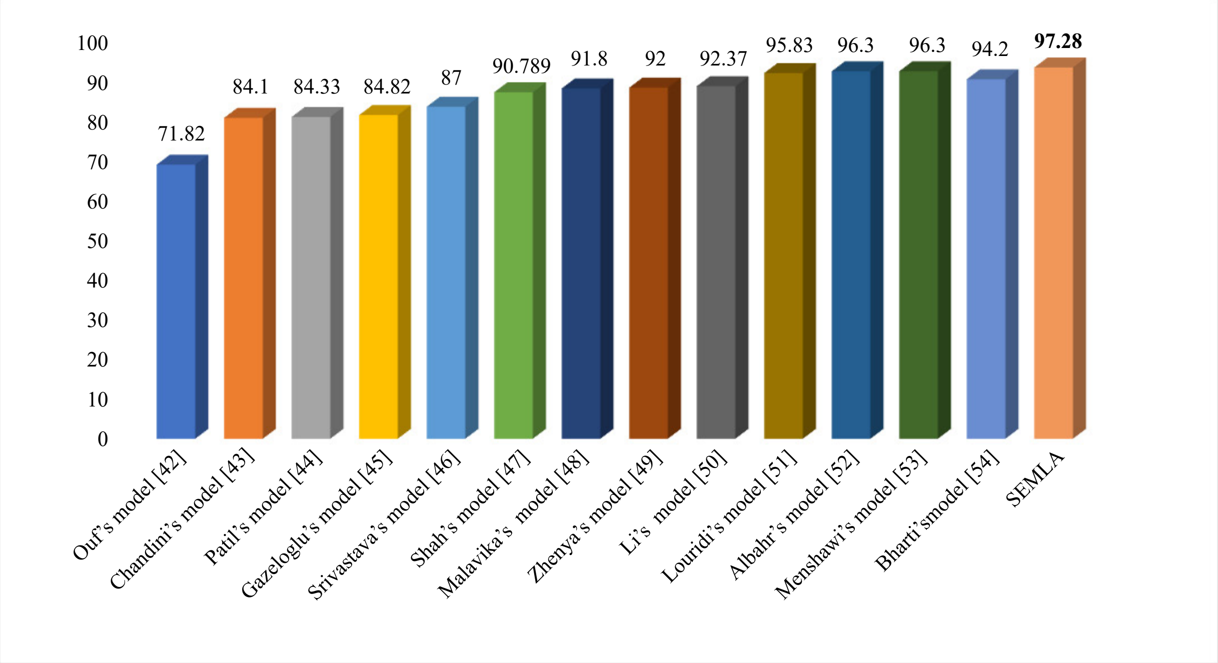


Figure 11. Comparison of SEMLA with other models in the literature

5. Conclusions

An ensemble model for heart disease prediction is presented in the proposed work. The accuracy, Dor, F_1_score, and log_loss of the proposed model's prediction to that of LR, RF, LDA, Nsvc, MLP classifiers have been evaluated. On data with all attributes, significant results are obtained as compared to the base models. The accuracy achieved is 97.28%, Dor of 1562.1, F_1_score of 0.973, roc-auc score of 0.992, 0.946 Mcc, and log_loss of 0.941. The learning models were then applied to data with only categorical variables. SEMLA has a value of 88.72% accuracy, 63.53 Dor, 0.887 F_1_score, 0.929 auc_score, 0.775 Mcc, and a value of 3.897 log_loss. Effective performance of SEMLA has been observed considering all features and categorical data with 16 ms and 10 ms execution times for prediction, respectively. The efficacy of the proposed model is compared to that of previous studies published between 2017 and 2021. We foresee significantly better outcomes if the suggested model (SEMLA) is indeed applied to large clinical data obtained through expert pathologists. The study's scope can be extended in the future to include other chronic conditions and industries.

Acknowledgments

We would like to thank Ms. Aafiya Ganachari for her valuable addition to the introduction section. The authors also appreciate anonymous reviewers for their valuable reviews, which have helped improve the technical writing of the paper.

Conflict of interest

The authors declare no competing interests.

References

[1] Yang H, Lin X, Li J, Zhai Y, Wu J. A review of mathematical models of COVID-19 transmission. *Contemporary Mathematics*. 2023; 4(1): 75-98. Available from: https://doi.org/10.37256/cm.4120232080.

[2] Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: A review. In: *Proceedings of Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA)*. Coimbatore, India: IEEE; 2018. p. 910-914. Available from: https://doi.org/10.1109/ICECA.2018.8474918.

[3] Park DJ, Park MW, Lee H, Kim Y-J, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific Reports*. 2021; 11: 7567. Available from: https://doi.org/10.1038/s41598-021-87171-5.

[4] Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. *IEEE Access*. 2019; 7: 180235-180243. Available from: https://doi.org/10.1109/ACCESS.2019.2952107.

[5] Centers for Disease Control and Prevention. *Underlying cause of death 1999-2020*. Available from: https://wonder.cdc.gov/ucd-icd10.html [Accessed 2nd February 2022].

[6] World Health Organization. *Cardiovascular diseases (CVDs)*. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 31st January 2022].

[7] Lapp D. *Heart disease dataset: Public health dataset*. kaggle. Available from: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset [Accessed 5th January 2022].

[8] Roccetti M, Delnevo G, Casini L, Silvia M. An alternative approach to dimension reduction for pareto distributed data: A case study. *Journal of Big Data*. 2021; 8: 39. Available from: https://doi.org/10.1186/s40537-021-00428-8.

[9] Kearney MW. Cramer's V. In: Allen M. (ed.) *The SAGE encyclopedia of communication research methods*. SAGE Publication; 2017. Available from: https://doi.org/10.4135/9781483381411.

[10] Ştefan R-M. A comparison of data classification methods. *Procedia Economics and Finance*. 2012; 3: 420-425. Available from: https://doi.org/10.1016/S2212-5671(12)00174-8.

[11] Çelik Ö. A research on machine learning methods and its applications. *Journal of Educational Technology and Online Learning*. 2018; 1(3): 25-40. Available from: https://doi.org/10.31681/jetol.457046.

[12] Sarker IH. Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science*. 2021; 2: 160. Available from: https://doi.org/10.1007/s42979-021-00592-x.

[13] Jurafsky D, Martin JH. Logistic regression. In: *Speech and language processing*. 3rd ed. Pearson; 2021. p.78-101.

[14] Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*. 2019; 7(4): 221-248. Available from: http://doi.org/10.1089/big.2018.0175.

[15] Ali Z, Shahzad SK, Shahzad W. Performance analysis of support vector machine based classifiers. *International Journal of Advanced and Applied Sciences*. 2018; 5(9): 33-38. Available from: https://doi.org/10.21833/ijaas.2018.09.007.

[16] Tharwat A, Gaber T, Ibrahim A, Hassanien AE. Linear discriminant analysis: A detailed tutorial. *AI Communications*. 2017; 30(2): 169-190. Available from: https://doi.org/10.3233/AIC-170729.

[17] Driss SB, Soua M, Kachouri R, Akil M. A comparison study between MLP and convolutional neural network models for character recognition. In: *SPIE Conference on Real-Time Image and Video Processing*. USA: SPIE 2017. Available from: http://doi.org/10.1117/12.2262589.

[18] Odegua R. An empirical study of ensemble techniques (bagging, boosting and stacking). In: *Proceedings of the Deep Learning IndabaX*. Nairobi, Kenya: IndabaX; 2019. p.25-31.

[19] Xue D, Zhou X, Li C, Yao Y, Rahaman MM, Zhang J, et al. An application of transfer learning and ensemble learning techniques for cervical histopathology image classification. *IEEE Access*. 2020; 8: 104603-104618. Available from: http://doi.org/10.1109/ACCESS.2020.2999816.

[20] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011; 12(85): 2825-2830.

[21] Tharwat A. Classification assessment methods. *Applied Computing and Informatics*. 2021; 17(1): 168-192. Available from: https://doi.org/10.1016/j.aci.2018.08.003.

[22] Ting KM. Confusion matrix. In: Sammut C, Webb GI. (eds.) *Encyclopedia of machine learning and data mining*. Boston, MA: Springer; 2019. p.209. Available from: https://doi.org/10.1007/978-0-387-30164-8_157.

[23] Gupta N, Rawal A, Narasimhan VL, Shiwani S. Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. *IOSR Journal of Computer Engineering*. 2013; 11(5): 70-73.

[24] Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. In: *NESUG Proceedings: Health Care and Life Sciences*. Baltimore, Maryland: NESUG; 2010.

[25] Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*. 2022; 12: 5979. Available from: https://doi.org/10.1038/s41598-022-09954-8.

[26] Shreffler J, Huecker MR. *Diagnostic testing accuracy: Sensitivity, specificity, predictive values, and likelihood ratios*. USA: StatPearls; 2022.

[27] Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: Calders T, Esposito F, Hüllermeier E, Meo R. (eds.) *Machine learning and knowledge discovery in databases: European Conference, ECML PKDD*. Berlin: Springer; 2014. Available from: https://doi.org/10.1007/978-3-662-44851-9_15.

[28] Wu S, Flach P. A scored AUC metric for classifier evaluation and selection. In: *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*. 2005.

[29] Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC curve. *International Journal of Epidemiology*. 2020; 49(4): 1397-1403. Available from: https://doi.org/10.1093/ije/dyz274.

[30] Aggarwal A., Kasiviswanathan SP, Xu Z, Feyisetan O, Teissier N. Label inference attacks from log-loss scores. *Proceedings of the 38th International Conference on Machine Learning, PMLR*. 2021; 139: 120-129.

[31] Vovk V. The fundamental nature of the log loss function. In: Beklemishev LD, Blass A, Dershowitz N, Finkbeiner B, Schulte W. (eds.) *Fields of logic and computation II*. Cham: Springer; 2015. p.307-318. Available from: https://doi.org/10.1007/978-3-319-23534-9_20.

[32] Bahadoran Z, Mirmiran P, Zadeh-Vakili A, Hosseinpanah F, Ghasemi A. The principles of biomedical scientific writing: Results. *International Journal of Endocrinology and Metabolism*. 2019; 17(2): e92113. Available from: https://doi.org/10.5812/ijem.92113.

[33] Vieira RF, de Lima RC, Mizubuti ESG. How to write the discussion section of a scientific article. *Acta Scientiarum. Agronomy*. 2019; 41(1): e42621. Available from: https://doi.org/10.4025/actasciagron.v41i1.42621.

[34] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: A survey. *International Journal of Engineering & Technology*. 2018; 7: 684-687. Available from: https://doi.org/10.14419/ijet.v7i2.8.10557.

[35] Jindal H, Agarwal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithm. *IOP Conference Series: Materials Science and Engineering*. 2020; 1022: 012072. Available from: https://doi.org/10.1088/1757-899X/1022/1/012072.

[36] Zhang R, Ma S, Shanahan L, Munroe J, Horn S, Speedie S. Automatic methods to extract New York heart association classification from clinical notes. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Kansas City, MO, USA: IEEE; 2017. Available from: https://doi.org/10.1109/BIBM.2017.8217848.

[37] Ma L-Y, Chen W-W, Gao R-L, Liu L-S, Zhu M-L, Wang Y-J, et al. China cardiovascular diseases report 2018: An updated summary. *Journal of Geriatric Cardiology.* 2020; 17(1): 1-8. Available from: https://doi.org/10.11909/j.issn.1671-5411.2020.01.001.

[38] Shrestha R, Chatterjee JM. Heart disease prediction system using machine learning. *LBEF Research Journal of Science, Technology and Management*. 2019; 1(2): 115-132.

[39] Nashif S, Raihan MR, Islam MR, Imam MH. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*. 2018; 6(4): 854-873. Available from: https://doi.org/10.4236/wjet.2018.64057.

[40] Ravikumar V, Bhavani M. Effective heart disease prediction using machine learning. *Journal of Engineering Sciences*. 2021; 12(12): 273-285.

[41] Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics*. 2020; 21: 278. Available from: https://doi.org/10.1186/s12859-020-03626-y.

[42] Ouf S, ElSeddawy AIB. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *Journal of Southwest Jiaotong University*. 2021; 56(4): 220-240. Available from: https://doi.org/10.35741/issn.0258-2724.56.4.19.

[43] Chandini R, Rao KV. Ensembling of SVM and decision tree for prediction of heart disease. *International Journal of Engineering Research & Technology*. 2021; 10(10): 298-301.

[44] Patil C, Dinesh, Patil DD, Subramanium P. Prediction of cardiovascular disease using machine learning techniques. *International Advanced Research Journal in Science, Engineering and Technology*. 2021; 8(11): 219-224.

[45] Gazeloglu C. Prediction of heart disease by classifying with feature selection and machine learning methods. *Progress in Nutrition*. 2020; 22(2): 660-667. Available from: https://doi.org/10.23751/pn.v22i2.9830.

[46] Srivastava K, Choubey DK. Heart disease prediction using machine learning and data mining. *International Journal of Recent Technology and Engineering*. 2020; 9(1): 212-219. Available from: https://doi.org/10.35940/ijrte.F9199.059120.

[47] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. *SN Computer Science*. 2020; 1: 345. Available from: https://doi.org/10.1007/s42979-020-00365-y.

[48] Malavika G, Rajathi N, Vanitha V, Parameswari P. Heart disease prediction using machine learning algorithms. *Bioscience Biotechnology Research Communications*. 2020; 13(11): 24-27. Available from: http://dx.doi.org/10.21786/bbrc/13.11/6.

[49] Zhenya Q, Zhang Z. A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Medical Informatics and Decision Making*. 2021; 21: 73. Available from: https://doi.org/10.1186/s12911-021-01436-7.

[50] Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A. Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*. 2020; 8: 107562-107582. Available from: https://doi.org/10.1109/ACCESS.2020.3001149.

[51] Louridi N, Douzi S, Ouahidi BE. Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*. 2021; 8: 133. Available from: https://doi.org/10.1186/s40537-021-00524-9.

[52] Albahr A, Albahar M, Thanoon M, Binsawad M. Computational learning model for prediction of heart disease using machine learning based on a new regularizer. *Computational Intelligence and Neuroscience.* 2021; 2021: 8628335. Available from: https://doi.org/10.1155/2021/8628335.

[53] Menshawi A, Hassan MM, Allheeib N, Fortino G. A hybrid generic framework for heart problem diagnosis based on a machine learning paradigm. *Sensors.* 2023; 23(3): 1392. Available from: https://doi.org/10.3390/s23031392.

[54] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande SD, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. *Computational Intelligence and Neuroscience*. 2021; 2021: 8387680. Available from: https://doi.org/10.1155/2021/8387680.