E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 03-07-2023
Accepted: 14-08-2023

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 1; Issue 1; 2023; Page No. 496-502

Construction of model for the classification of chronic disease and particularly heart

[1]Prathima Y and [2]Dr. Manish Saxena

[1]Research Scholar, Department of Computer Science, Himalayan University, Arunachal Pradesh, India
[2]Assistant Professor, Department of Computer Science, Himalayan University, Arunachal Pradesh, India

DOI: https://doi.org/10.5281/zenodo.12820822

Corresponding Author: Prathima Y

Abstract

Medical services provide gigantic information on every day ground having diverse structures like printed, images, numbers pool and so forth. However, there is absence of devices accessible in healthcare to process this data. Data mining frame works are utilized to extricate information from this data which can be utilized by media proficient individual to figure future procedures. Heart illness is the primary driver of death in the masses. Early recognizing and hazard expectations are essential for patient's medicines and specialists' analysis. Data mining has found success in highly visible industries such as retail marketing and e-commerce, which has led to its application in the healthcare context. Predictive analysis and processing can be helpful in helping patients determine the source of their illness, especially with the increasing demand for medical data. Excessive processing of cardiovascular disease-related medical data has led to growth in a certain order that limits manual analysis for parameter prediction in decision-making. Advancements in medical diagnosis systems have demonstrated the advantages of computer algorithms. In recent years, there has been a decrease in the number of deaths. It is also proven that deaths from serious, life-threatening conditions like heart disease are declining. Benefits are obtained by accurate diagnosis and early identification of medical conditions through patient data analysis. However, to have a more accurate and timely analysis, the usual algorithms need to be enhanced. Predictive outcomes from the analysis are required for diagnosis. As a result, the death rates can be further decreased. Many research efforts are undertaken to enhance the outcomes; yet, there is room for improvement in the current methodologies.

Keywords: Data mining, life-threatening, hazard, medicines, e-commerce, methodologies

Introduction

Data mining has found success in highly visible industries such as retail marketing and e-commerce, which has led to its application in the healthcare context. Predictive analysis and processing can be helpful in helping patients determine the source of their illness, especially with the increasing demand for medical data. Excessive processing of cardiovascular disease-related medical data has led to growth in a certain order that limits manual analysis for parameter prediction in decision-making.

Advancements in medical diagnosis systems have demonstrated the advantages of computer algorithms. In recent years, there has been a decrease in the number of deaths. It is also proven that deaths from serious, life-threatening conditions like heart disease are declining. Benefits are obtained by accurate diagnosis and early identification of medical conditions through patient data analysis. However, to have a more accurate and timely analysis, the usual algorithms need to be enhanced. Predictive outcomes from the analysis are required for diagnosis. As a result, the death rates can be further decreased. Many research efforts are undertaken to enhance the outcomes; yet, there is room for improvement in the current methodologies.

For disease prediction, individualized treatment plans, disease outbreak prediction, medication development, drug interactions, and other predictive services, the medical industry uses machine learning algorithms and methodologies. In order to raise the standard of care, the healthcare sector began to apply machine learning algorithms in an efficient manner using a wide range of data types. Machine learning-based decision support systems have revolutionized the way doctors identify and treat patients (Nazari *et al*., 2018) [15]. In the healthcare sector, the application of machine learning algorithms has changed how decisions are made, particularly in the areas of disease

detection and predictive services (Yan *et al*., 2006) [16]. Decision-making including the use of healthcare data has become essential, ranging from robotic surgery to the discovery of novel drugs. The healthcare business has undergone a transformation thanks to the daily generation of thousands of medical data and the exploration and use of this volume of data to uncover hidden information through machine learning. Machine learning methods can be used with a variety of data kinds, including text, numbers, pictures, signals, multimedia files, and photographs.

There are two steps in the healthcare process: (i) investigation and examination; and (ii) treatment and monitoring. By extracting pertinent facts from the data, machine learning models challenge preconceived notions about how a process will turn out. It gets the ability to make decisions that can help patients, management, and the healthcare system by using the patterns found in the data.

Feature selection techniques enable the use of supervised learning to comprehend information and choose aspects that are important for disease prognosis. Numerous studies have examined features and their significance, and they have come to the conclusion that feature selection is crucial to enhancing disease diagnosis (Shankar *et al*. 2018; Nazari *et al*. 2018) [3, 18]. Enhancing classification accuracy and lowering the complexity of developing and training on huge datasets are two benefits of feature selection. Several studies employ feature selection to increase learning rates by reducing the impact of redundant and noisy features, which lowers the classifiers' learning rate.

Role of feature selection in data mining

Feature selection is the process of selecting feature subsets that are relevant to the classes. The selected subset contains maximum information about the target classes. Feature selection reduces the dimensions of the dataset and plays a vital role in dimension reduction. In disease diagnosis, features are the symptoms and indicators of the illness severity, understanding characteristics can help to comprehend the full process and functions of any disease. In genetic studies feature selection helps to distinguish between genes that cause disease and normal cells. Feature selection and dimension reduction slightly differ from each other, where feature selection picks and remove features in a subset while dimension reduction tries to produce combinations of attributes.

Features are best to identify between class levels. The presence and absence of disease can be discriminated using a single feature out of all possible features contained in the dataset. Identifying potential features is the purpose of any feature selection algorithms. Large Features in a dataset can generate time complexity and memory concerns in model building and presence of huge number of features may induce overfitting in the models. Features selection helps to improve the interpretability of the system when features are constrained.

Feature selection

The process of choosing pertinent characteristics and eliminating unnecessary features from a dataset is known as feature selection. In a dataset, samples are represented by rows and features by columns. A row in a heart disease dataset represents a patient record, while columns represent features. Any feature selection technique should aim to enhance the model's performance in addition to choosing the best features. Three different feature selection techniques exist: the filter approach, the embedding method, and the wrapper method. To choose the optimal subset of traits, each of these techniques operates in a unique way. Gene research (Liu *et al*. 2014) [12], image analysis, intrusion detection, defect diagnosis, and other fields all make extensive use of feature selection techniques.

Filter Method: To assess a feature's utility, the filter method filters the features using statistical approaches (Li *et al*. 2014) [12] and assessment methodologies. Information theory and the distance function are used by the assessment methods to rank the features. Based on the highest ranks, the ranked features are sorted, ordered, and chosen. Filter techniques don't require models and are computationally cheap. The features are ranked using statistical methods such mutual information, chi-square, and correlation.

The Wrapper Method is a technique that chooses features based on learning models. A predictive model is utilized for evaluation and a search strategy is employed to choose the attributes. According to the model's performance, the features are chosen. Forward selection, backward selection, and backward elimination techniques are all used in the selection process. Only those features are chosen for the subset that positively impact the model's performance; other features are removed. In addition to choosing the pertinent features, this kind of feature selection enhances the performance of the model. When compared to filter methods, wrapper methods are more effective.

The embedded technique is a kind of nested strategy in which the features are chosen as a means of assisting in the training of the learning process, and the selection process is directed as a search process. According to Xiao *et al*. (2008) [18], the features are chosen throughout the model construction process' training phase. In contrast to wrapper techniques, this approach is more reliable, prevents overfitting, and doesn't require a separate training set to evaluate the features. The embedded technique is not transferable to other classifiers; it depends on the kind of classifier that is used to assess the features. Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Random Forest, and LDA are among examples.

Proposed methodology

The practice of selecting a subset of features with the goal of enhancing performance and classification outcomes is known as feature selection. There are several uses for feature selection, including data reduction, time complexity reduction, accuracy and performance improvement, and more. The term "feature selection" usually refers to the process of choosing the greatest and most ideal features for the target classes. This study's primary goal is to identify the most pertinent features for a given dataset while removing any unnecessary features to increase classification accuracy.

Figure-1 displays the suggested model's general layout. Features that are highly informative are chosen for classification in order to lower the classification error rate, while irrelevant features frequently increase error rates.

FS generates a Y-dimensional space of SY for an X dataset with Y features, where a subspace i from SYi contains the

most pertinent features that describe the target class. The methods of filter, embedding, and wrapper are included in the common feature selection. Wrapper approaches employ algorithms for rating criteria and greedy search to optimize learning repeatedly. The grading criteria make sure that the choices are consistent with the output of the algorithms. In the feature space SY, the search method creates a permutation of distinct features Y. Subset production is halted by applying halting criterion C. The best feature subset is chosen based on greater performance by using the nearest neighbor as the basis classifier.

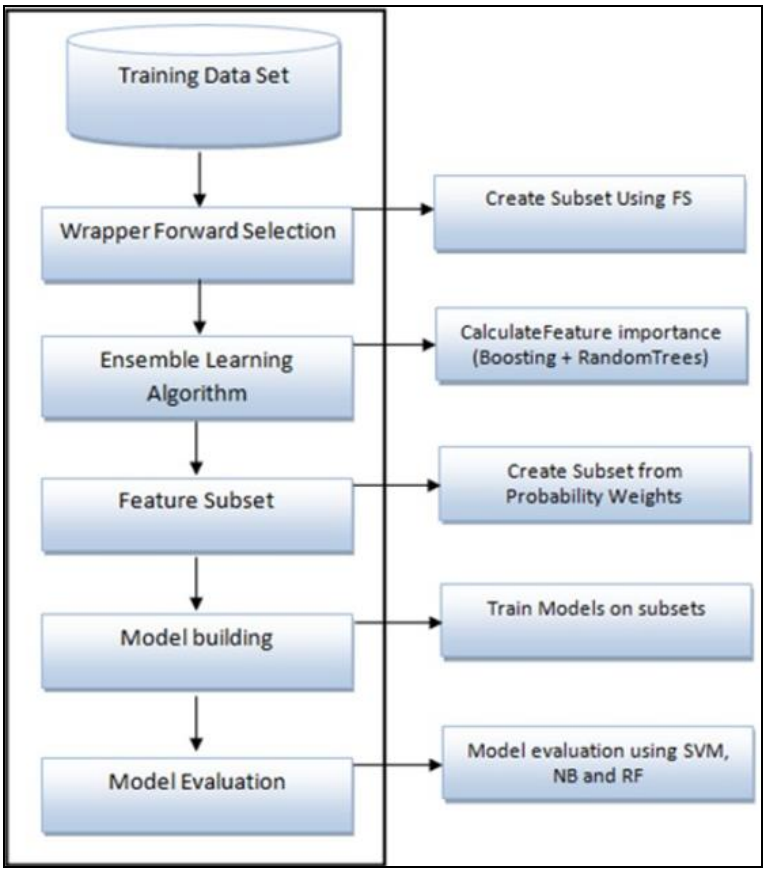

Fig 1: Schematic representation of proposed model

To achieve generality and further prevent strong correlation between the variables, random trees are constructed from the resulting subset. Table-1 lists the names and symbols that were utilized.

Table 1: List of symbols

Symbol	Description
$X= (x_1,x_2..., x_d)$	X is a feature vector
Y	y is the target class
E	Represents the ensemble of decision tree classifiers
P_k	Hyper parameters of a classifier
$X_{sub} = ((X_1,y_1)...,(X_n, y_n))$	Subset vector of features
CR (X, C) = v(X,C1) 1 Rk	Class relationship score of a feature
Ci = mvote{CRk}k 1	Majority vote of random trees
R_k	Random Trees
O_k	Subset with k features
b+	A feature that contributes to classifier performance
C	Stopping criterion
SY	Y in feature subspace

Algorithm 1

Input: $X=(x_1,x_2,...,x_d)$ (the whole dimensional space with SY)

Output: $O_k = \{b_i \mid i=1, 2,..., k; b_i \in X\}$ where k = (0, 1, 2,..., SY) O_k returns a subset with k features,

Where k < SY

Initialize selection: O_k = empty, k = 0

(k is the feature size) Step 1:

$b+= \arg\max I(b_k+ b)$, where $b \in X-O_k$

$O_k +1= O_k + b+$

k = k + 1

Go to Step 1

b+ is the feature that adds to the performance of the classifier and added to O_k repeat until C is reached

Go to 1

Ensemble Learning

Based on classification trees, the most potent learning ensemble tree model constructs trees using two random steps. Using bootstrap samples, trees are grown on each individual feature in the first stage. Features and the matching members of the intermediary nodes make up the leaf node. A new tree is added to the bottom of the current trees in the second step. For the target class, each tree represents a sort of class preference. The target class is assigned to the trees with the highest class preferences.

Several bootstrapped samples Bt from the training set are

combined to create distinct trees for each variable, which are then used to build the ensemble classifier using random classification trees. Assume that X = (x1,x2,...,xd) is a feature vector. The input vector X contains random variables for each feature (x1). Let y represent the disease categorization, where y = 0 denotes absence and y = 1 denotes presence. Now, using ensembles of the classifiers E such that E = (E1(X),...,Ek(X)), features from X predict y. Each E1(X) is a decision tree, and Pk=(Pk1,...,Pkn) represents the hyperparameters of a decision tree for the classifier E1(X). The formula for the decision tree classifier is Ek(X) = E(X|Pk). In the input vector X, class y receives the most votes from each tree based on hyperparameter Pk.
Based on the hyper parameters, Pk calculates the subset of trees Xsub; the vector of the subset is then indicated by Xsub = ((X1,y1)...,(Xn, yn)). The occurrence of C1 for the classifier Ek($1 \leq k \leq K$) determines the probability of class C1 for the ensemble Ek(X). Votes and the total number of trees in the forest are used to calculate the class relationship score, or CR. It is depicted as

$$CR(X,C1) = \frac{y(X,C1)}{RK}$$

Sr = Σz(yi, C(Xi)) × CR(X,C1)) can be used to get the feature selection score (where z = 0 for i ≔ j and 1 otherwise) and C(X) is the class label that the classifier E(X|P1) assigned. Among several trees in Rk, a tree's majority vote is determined by

$$C^{i} = \text{mvote}\{C^{Rk}\}k_1.$$

The feature importance, which is based on the variation between the original samples and the out-of-bag samples, must be calculated next. The mean error values of the permuted and original samples are used to compute the difference between the samples. Gini impurity is used to compute the mean error. Gini (n) = 1- ∑2 $fYi2$, where fi is the frequency of class Y in the node n, is the formula for calculating the Gini index of the tree with node n. The feature relevance is shown in the larger decrease in error. The association can be found by comparing the mean errors of the original and permutated samples; mean error is used to exclude weak associations. The mean error may be found using

$$E^{oob} = \frac{1}{N_{oob}} \sum_{i=1}^{n} (Y, Y'^{oob}).$$

The feature importance score from the previous technique is used in the suggested feature ranking strategy, and a new random forest is built utilizing the feature importance score as a weighting requirement for feature sampling. The best split from the feature sampling is chosen using the features' probability Pi, which is proportional to weight wi. The following formula provides the weighting criterion wi:

$$wi = \frac{1}{Pi}.$$

By removing noisy features from the dataset, the suggested model increases the accuracy of classification through feature sampling based on the variable importance scores.

Algorithm 2
Input: X=(x1,x2,...,xd) (the whole dimensional space with SY),
Rk = {bi | i=1, 2,..., k; bi ∈ X}
where k = (0, 1, 2,..., SY)
K = Tree Numbers Output: Random Trees Rk For k to 1 → k do

Build a bagged subset of samples from Rk.
Select randomly X features. For X to 1 → SY do
Calculate decrease in the leaf node impurity.
Calculate the feature importance score
Resample using feature weights
Build random forest and split trees based on weighting criteria

While features that grow into trees have a lot of variance, those filtered through forward selection have significantly less variation because to minimal nearest neighbor k. To prevent overfitting and lower variance, the training data might be divided into many bootstrapped samples. The variation in the trees can be decreased and overfitting can be prevented by averaging the various samples. The samples created at the random split would eventually have two separate variables at the very least when building trees using different bootstrapped samples. Each variable becomes independent as a result, and the correlation between the trees declines. Small training instance numbers are typically when the variance problem occurs, and this can be bootstrapped to prevent larger variances. Additionally, as the number of training instances rises, variation naturally falls. By averaging the bootstrapped samples, the variance issue can be lessened. The classification performance is enhanced and the variance is decreased by the averaging. Fifteen distinct datasets are used to examine several strategies, including FSNBb, FSSVMb, GASVMb, GANBb, GASVMb, and GARFb, in order to assess the effectiveness of the suggested ensemble learning method, ESFS.

Experiments and analysis
Dataset
Fifteen distinct datasets are selected from the UCI repository and subjected to the suggested feature selection procedure. Table-2 lists the feature numbers that relate to the datasets. Several datasets with different numbers of rows, columns, feature dimensions, feature types, and class levels were used for this study in order to assess the model's performance. With 8124 data instances, the Mushroom dataset has the most; Audiology has the most features (69); the Zoo dataset has the fewest data instances (101); and the Diabetes dataset has the fewest features (9). To capture the model's performance across all dimensions and examine the model's behavior with respect to data size, feature numbers, and various scales, features with varying data types, sizes, scales, and numbers are helpful. For this study, benchmark datasets and datasets that are well-liked by the data mining community were selected. The research community

generally favors the fifteen distinct datasets selected for the study for investigations into dimensionality reduction, subset selection, feature selection, feature extraction, and other data mining issues.

Table 2: Datasets used in the study

S. No	Dataset	Instances	No of Features
1	Mushroom	8124	22
2	Thyroid	7200	22
3	Diabetes	768	9
4	Liver	584	11
5	Breast Cancer	569	32
6	Heart (SA)	462	10
7	CKD	400	24
8	Dermatology	366	34
9	Ionosphere	351	34
10	Tumor Data	339	17
11	Heart (Cleveland)	303	14
12	Heart (Statlog)	270	13
13	Audiology	226	69
14	Lymphography	148	18
15	Zoo	101	17

Evaluation Metrics

The performance of the classifiers is displayed through true positive, true negative, false positive, and false negative ratios using a confusion matrix. Genuine positive (TP) indicates when a positive prediction is correct; false positive (FP) indicates when a positive prediction is incorrect; true negative (TN) indicates when a negative prediction is correct; and false negative (FN) indicates when a negative prediction is incorrect. In Table-3, the confusion matrix is displayed. We can find accuracy using the following formula:

$$\text{Accuracy} = (T\,P + T\,N) / (T\,P + T\,N + F\,P + F\,N),$$

whereas sensitivity, specificity, precision, and F-score is given by:

$$\text{Sensitivity} = (T\,P) / (T\,P + F\,N),$$

$$\text{Specificity} = (T\,N) / (T\,N + F\,P), \quad \text{Precision} = (T\,P) / (T\,P + F\,P),$$

$$\text{F-score} = (2 * T\,P) / (2 * T\,P + F\,P + F\,N).$$

Python is used in the development of the suggested feature selection algorithm and classifiers. Using ensemble feature selection techniques like FSNBb, FSSVMb, GASVMb, GANBb, and GARFb, the suggested method's performance is compared. To assess the suggested approach, the classification accuracy on the training set of the chosen datasets is contrasted. Table-3 provides a description of the specifics of the feature selection techniques together with their acronyms.

Table 3: Models used in the study

S. No	Algorithm	Abbreviation
1	SFS + Bagging+ NB	FSNBb
2	SFS + Bagging+ SVM	FSSVMb
3	GA + Bagging+ NB	GANBb
4	GA + Bagging+ SVM	GASVMb
5	GA + Bagging+ RF	GARFb
6	SFS + Bagging+ RF	ELFS* (Proposed method)

Model Building

Forward selection, genetic search, and bagging techniques are used to create various ensemble approaches based on Naive Bayes, SVM, and RF in order to test the suggested model. 40% of the dataset is designated as a testing set and 60% as a training set. The built-in models undergo training on the training dataset, and they undergo evaluation on the testing dataset.

Results and Discussion

There are two parts in the performance evaluation process. Firstly, all features from the Thyroid, Diabetes, Liver, Heart (SA), Tumor, Heart (Cleveland), Heart (Statlog), Lymphography, and Zoo datasets are used to evaluate the models. The model's performance is assessed and feature selection is used in the second stage. Based on the findings of the evaluation of all features and certain features, the performance is interpreted. Table-4 lists the classification models' accuracy when all features are used on all datasets.

Table 4: Accuracy of ensemble algorithms on datasets using all features

Dataset	Features	FSNBb	FSSVMb	GASVMb	GANBb	GARFb	ELFS
Mushroom	22	0.66	0.79	0.83	0.7	0.86	0.81
Thyroid	22	0.73	0.84	0.95	0.71	0.96	0.83
Diabetes	9	0.79	0.76	0.94	0.91	0.95	0.96
Liver	11	0.80	0.85	0.93	0.74	0.97	0.91
Breast Cancer	32	0.84	0.96	0.91	0.73	0.81	0.86
Heart (SA)	10	0.79	0.82	0.94	0.71	0.95	0.89
CKD	24	0.59	0.78	0.86	0.79	0.79	0.95
Dermatalogy	34	0.75	0.81	0.79	0.52	0.83	0.8
Ionosphere	34	0.6	0.77	0.81	0.7	0.88	0.90
Tumor Data	17	0.57	0.67	0.73	0.76	0.78	0.81
Heart (Cleveland)	14	0.68	0.79	0.85	0.76	0.89	0.91
Heart (Statlog)	13	0.74	0.81	0.87	0.69	0.91	0.92
Audiology	69	0.63	0.74	0.73	0.82	0.90	0.88
Lymphography	18	0.75	0.73	0.88	0.86	0.82	0.79
Zoo	17	0.67	0.78	0.87	0.73	0.86	0.86
Average		0.71	0.79	0.86	0.74	0.88	0.87

Classification models GASVMb (83%) and GARFb (86%) had a high accuracy for 22 characteristics in the Mushroom dataset. GASVMb (95%) and GARFb (96%) obtained an accuracy rate greater than 90% for the Thyroid dataset. Models GASVMb (94%), GANBb (91%), GARFb (95%) and ELFS (96%) received greater accuracy scores of more than 90% for the diabetes dataset compared to other models. GARFb achieves a greater accuracy of 97% for the liver dataset. The dataset on breast cancer showed that FSSVMb had the maximum accuracy of 96%. GARFb attained a greater accuracy of 95% for the heart (SA) dataset. The suggested strategy yielded a greater accuracy of 95% for the CKD dataset. With regard to the dermatological dataset, FSSVMb's accuracy was greater at 81%. Compared to previous models, the suggested model ELFS achieved higher accuracy: 90% on the Ionosphere dataset, 81% on the Tumour dataset, 91% on the Cleavland dataset, and 92% on the Statlog dataset. While GASVMb achieved a greater accuracy of 88% on Lymphography and 87% on Zoo dataset, GARFb reached an accuracy of 90% on Audiology dataset. In comparison to other models, the suggested model

produced greater accuracy across six datasets. As indicated in Table-5, the suggested model ELFS achieved a high accuracy rate on eleven datasets by using feature selection for classification.

Table 5: Accuracy of ensemble algorithms using features selection

Dataset	Features	FSNBb	FSSVMb	GASVMb	GANBb	GARFb	ELFS
Mushroom	12	0.86	0.89	0.93	0.90	0.96	0.87
Thyroid	6	0.83	0.82	0.85	0.82	0.9	0.93
Diabetes	5	0.79	0.76	0.94	0.91	0.95	0.96
Liver	7	0.81	0.83	0.89	0.84	0.93	0.94
Breast Cancer	11	0.78	0.86	0.89	0.83	0.91	0.96
Heart (SA)	6	0.75	0.72	0.88	0.81	0.89	0.93
CKD	14	0.69	0.75	0.76	0.74	0.89	0.85
Dermatology	15	0.78	0.8	0.83	0.82	0.93	0.91
Ionosphere	13	0.65	0.81	0.86	0.71	0.78	0.85
Tumor Data	9	0.59	0.77	0.84	0.67	0.87	0.94
Heart (Cleveland)	7	0.68	0.79	0.85	0.76	0.89	0.91
Heart (Statlog)	6	0.79	0.8	0.86	0.73	0.95	0.96
Audiology	19	0.57	0.64	0.79	0.77	0.94	0.89
Lymphography	8	0.65	0.78	0.84	0.81	0.87	0.90
Zoo	9	0.95	0.91	0.96	0.93	0.96	0.97
Average		0.74	0.80	0.86	0.8	0.91	0.92

In the medical field, it is more expensive to anticipate a positive case as negative than a positive case as positive. The purpose of testing for the presence of a disease is to rule it out, thus diagnostic and treatment expenses are squandered if a patient is expected to be false positive. In the case of heart illness, a patient's life expectancy may worsen if they are determined to be falsely negative; conversely, if a patient is expected to have heart disease and is therefore falsely positive, the additional expenses incurred for the clinical diagnosis would be squandered. The ratio of false negatives to false positives is crucial to classification performance. To assess the model's performance in terms of false positives and false negatives, one can compare the accuracy score. The model performs better the higher the precision rate. The precision values of the models with all features and with the feature selection approach applied are displayed in Tables-6 and 7, respectively.

Table 6: Precision score of ensemble models using all features

Dataset	Features	FSNBb	FSSVMb	GASVMb	GANBb	GARFb	ELFS
Mushroom	22	0.64	0.76	0.81	0.65	0.85	0.75
Thyroid	22	0.74	0.84	0.95	0.96	0.96	0.83
Diabetes	9	0.86	0.94	0.96	0.94	0.98	0.98
Liver	11	0.89	0.96	0.97	0.77	0.99	0.93
Breast Cancer	32	0.85	0.99	0.99	0.85	0.81	0.85
Heart (SA)	10	0.91	0.89	0.96	0.64	0.96	0.91
CKD	24	0.90	0.90	0.93	0.87	0.70	0.93
Dataset	Features	FSNBb	FSSVMb	GASVMb	GANBb	GARFb	ELFS
Dermatology	34	0.76	0.87	0.88	0.78	0.80	0.63
Ionosphere	34	0.53	0.78	0.87	0.75	0.89	0.92
Tumor Data	17	0.67	0.68	0.76	0.78	0.82	0.24
Heart (Cleveland)	14	0.65	0.82	0.92	0.77	0.96	0.97
Heart (Statlog)	13	0.80	0.81	0.91	0.64	0.93	0.93
Audiology	69	0.60	0.73	0.72	0.82	0.92	0.94
Lymphography	18	0.80	0.78	0.91	0.91	0.80	0.82
Zoo	17	0.76	0.87	0.16	0.63	0.73	0.83
Average		0.76	0.84	0.85	0.78	0.87	0.83

Table 7: Precision score of ensemble models using features selection

Dataset	Features	FSNBb	FSSVMb	GASVMb	GANBb	GARFb	ELFS
Mushroom	12	0.95	0.97	0.97	0.97	0.97	0.93
Thyroid	6	0.83	0.82	0.85	0.82	0.91	0.94
Diabetes	5	0.8	0.73	0.72	0.96	0.92	0.97
Liver	7	0.87	0.88	0.92	0.87	0.95	0.97
Breast Cancer	11	0.93	0.95	.90	0.90	0.96	0.97
Heart (SA)	6	0.69	0.68	0.91	0.86	0.91	0.95
CKD	14	0.8	0.89	0.9	0.9	0.89	0.88
Dermatology	15	0.66	0.72	0.70	0.68	0.89	0.88
Ionosphere	13	0.9	0.9	0.88	0.76	0.85	0.89

Tumor Data	9	0.69	0.8	0.85	0.65	0.88	0.96
Heart (Cleveland)	7	0.65	0.79	0.91	0.70	0.86	0.9
Heart (Statlog)	6	0.84	0.8	0.9	0.77	0.95	0.97
Audiology	19	0.5	0.61	0.79	0.79	0.97	0.93
Lymphography	8	0.63	0.86	0.85	0.79	0.91	0.91
Zoo	9	0.91	0.91	0.96	0.89	0.96	0.98
Average		0.78	0.82	0.87	0.82	0.92	0.94

Conclusion

Heart Disease is a fatal disease by its nature. This disease makes a life-threatening complexity, for example heart attack and death. The significance of Data Mining in the Medical Domain is acknowledged and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied. A New technique of predicting heart disease is developed which resulted in better accuracy than the existing works. In this work adaboost algorithm has the greater accuracy than logistic regression for training data and ensemble have comparatively less accuracy than both adaboost and logistic regression. This project can be developed further by using the concept of Internet of things where sensors can be arranged in the vehicles of people who are probable of facing a heart disease and alerts can be sent faster to the nearest medical facilities. The research findings and the study purpose are related in the following ways. The suggested ensemble-based enhanced tree model is used to investigate the prediction model for categorizing the presence and absence of disease. The suggested ensemble tree model categorizes the illness states and chooses pertinent attributes. The suggested model's results demonstrate that feature selection enhances the model's functionality. The objective, which is to "diagnose the presence of chronic disease, mainly heart disease through proposed feature selection method," is satisfied by the proposed ensemble model's findings, which demonstrate that the presence or absence of heart disease may be predicted by choosing pertinent features.

References

1. Voruganti S. Effective IOT techniques to monitor the levels of garbage in smart dustbins. Int Res J Eng Technol. 2020;7(6):6549-6554.
2. Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iran J Basic Med Sci. 2016;19(5):476-482.
3. Surya Samantha B, Truth M, Sairam U. A Review on Using Crow Search Algorithm in Solving the Problems of Constrained Optimization. Res Rev Int J Multidiscip. 2018;4(2):1004-1009.
4. Bhanu Prakash MV, Sairam U. Feature Prospect of the VAST Applications of Machine Learning. Res Rev Int J Multidiscip. 2019;4(4):1265-1271.
5. Sairam U. Multi-Functional Blind Stick for Visually Impaired People. IEEE Explore; c2020. ISBN:978-1-7281-5371-1.
6. Sairam U, Bhanu Prakash MV. DL And ML Approaches Along with Blockchain Towards IoT Security. Int J Adv Sci Technol. 2020;29(4s):826-832.
7. Guru N, Dahiya A, Rajpal N. Decision Support System for Heart Disease Diagnosis Using Neural Network. Delhi Bus Rev. 2007;8(1):99-101.
8. Peters RM, Shanies SA, Peters JC. Fuzzy cluster analysis of positive stress tests, a new method of combining exercise test variables to predict extent of coronary artery disease. Am J Cardiol. 1995;76(10):648-651.
9. Chitra R, Seenivasagam V. Heart Attack Prediction System Using Fuzzy C Means Classifier. IOSR J Comput Eng. 2013;14(2):23-31.
10. Voruganti S. Local Security Enhancement and Intrusion Prevention in Android Devices. Int Res J Eng Technol. 2020;7(1):205-211.
11. Kapil S, Ch M, Ansari MD. On K-means Data Clustering Algorithm with Genetic Algorithm. In: Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE; c2016. p. 202-206.
12. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. In: Data Classification: Algorithms and Applications. CRC Press; c2014.
13. Kunta V, Tuniki C, Sairam U. Multi-Functional Blind Stick for Visually Impaired People. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES); c2020. p. 895-899.
14. Adepu Y, Boga VR, U S. Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition. In: 2020 IEEE International Conference for Innovation in Technology (INOCON); c2020. p. 1-5.
15. Nazari F, Noruzoliaee M, Mohammadian AK. Shared versus private mobility: Modeling public interest in autonomous vehicles accounting for latent attitudes. Transportation Research Part C: Emerging Technologies. 2018;97:456-477.
16. Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede MA, *et al*. The wheat and barley vernalization gene VRN3 is an orthologue of FT. Proceedings of the National Academy of Sciences. 2006;103(51):19581-19586.
17. Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsky P, Loibl S, *et al*. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Annals of Oncology. 2017;28(8):1700-1712.
18. Chen JH, Jang C, Xiao S, Ishigami M, Fuhrer MS. Intrinsic and extrinsic performance limits of graphene devices on SiO2. Nature nanotechnology. 2008;3(4):206-209.