Role of Feature Selection in Building High Performance Heart Disease Prediction Systems

Ekta Maini [1], Bondu Vankateswarlu[2], Arbind Gupta[3]

[1]Research Scholar,
Dayananda Sagar University Bengaluru,
ekta.marwaha@gmail.com

[2]Associate Professor
Dayananda Sagar University Bengaluru,
bonduvenkat-cse@dsu.edu.in

[3]Professor
Dayananda Sagar College of Engineering,
arbind.gupta@gmail.com

Abstract*: *In the last few years, there has been a tremendous rise in number of deaths due to heart diseases all over the world. In low and middle-income countries, heart diseases are usually not detected in early stages which makes the treatment difficult. Early diagnosis can help significantly in preventing these diseases. Machine learning based prediction systems offer a cost effective and efficient way to diagnose these diseases in an early stage. Research is being carried out to increase the performance of these systems. Redundant and irrelevant features in the medical dataset deteriorate the performance of prediction systems. In this paper, an exhaustive study has been done to improve the performance of the prediction systems by applying 4 feature selection algorithms. Experimental results prove that the use of feature selection algorithms provides a substantial increase in accuracy and speed of execution of the prediction system. The prediction system proposed in this study shall prove to be a great help to prevent heart diseases by enabling the medical practitioners to detect heart diseases in early stages.

***Keywords*: Machine Learning, feature selection, heart disease prediction**

(Article history: Received: 21st August 2019 and accepted 29th November 2019)

I. Introduction

There is a tremendous increase in the number of deaths because of cardiac ailments in the recent years. It is observed that heart diseases are not confined to a particular region but are rising alarmingly globally. These diseases have taken a huge toll of life particularly in developing countries like India [1]. Healthcare services in India have a great scope of improvement. Currently, India is positioned at 145th place among all the nations of world on the basis of quality and accessibility of healthcare [2]. There is a great need to provide cost effective and easily accessible tools which can aid in detecting the diseases in early stages. This is an age of data explosion and healthcare industry generate a huge volume of data in the form of Electronic Health Records (EHR), physician's prescription, data related to pharmacy and health insurance etc. A number of computationally powerful machine learning algorithms like Logistic regression, Decision tree, Naïve Bayes, Random Forest, Adaboost, Support Vector Machine have proved their competence in generating valuable information from the data obtained from healthcare data [3]. Risk of getting affected by a disease can be easily predicted using such information [4,5].These models are later hosted on the cloud so that these prediction models can be easily used by the medical practitioners to forecast the chances of heart diseases in the patients [6] .Thus Artificial intelligence can play a vital role in tackling the serious issues of quality, accessibility and affordability in healthcare in India. Though much research has been done in building such prediction models, yet there is a considerable scope of improvement of these systems.

It is observed that healthcare data may contain missing ,redundant or irrelevant attributes. The accuracy of the prediction system is tarnished due to the presence of such features [7]. Research suggests that use of feature selection algorithms to remove of such features may help to improve the accuracy of the prediction systems for various diseases [8,9]. Nalband et al proposed a system to predict knee joint disorders .The system was designed using apriori algorithm to select significant features along with Random Forest classifier [10]. Mohammad Shafenoor Ali et al proposed a brute force algorithm for determining important features in heart disease dataset. The study achieved an accuracy of 87% for heart disease prediction system [11]. Muhammad Usman et al applied Cuckoo search algorithms and Cuckoo optimization algorithms to eliminate the redundant features from heart disease dataset. Experimental outcomes indicate that the accuracy of the system increased to 87% [12]. Backward search method of feature selection was used by Balakrishnan S et al to identify the important features on diabetes II dataset. Results of the study suggest a considerable improvement in the performance of the system [13].

Through this research we aim to analyze the competence of various feature selection algorithms in improving the performance of heart disease prediction system. Characteristics, merits and demerits of filter, wrapper and embedded feature selection algorithms have been discussed in this paper. Experiments were carried out on Cleveland heart disease dataset provided by UCI data repository [14]. Least absolute shrinkage and selection operator (LASSO), minimum redundancy and maximum relevance (MRMR), genetic algorithm and Relief filter techniques of feature selection were used to remove the insignificant features from the dataset. The relative merits and demerits of these algorithms have been discussed in this paper. 5 machine learning techniques namely Logistic Regression, Random Forest, k-NN, Naïve Bayes, Support Vector Machine were used to develop prediction models. Specificity, accuracy and sensitivity and processing speed were assessed to analyze the performance of the system. This study contributes significantly by highlighting how the competence of AI based prediction system improves with the application of feature selection algorithms. The results obtained in this study also provide a deep insight regarding which feature selection algorithm works well with which classifier.

II. Feature Selection Algorithms

In the present era, voluminous amount of data is produced by the healthcare sector . Many medical diagnostic tests are done in hospitals to check if the patient suffers from some kind of heart disease or not. With the growth of artificial intelligence ,machine learning based prediction models have been built through which chances of heart disease in a patient can be forecast in an early stage. However, since the medical dataset has large number of attributes, the researcher faces a challenge to decide which attributes should be selected and which attributes which should be discarded for building a prediction model.. The dataset may consist of irrelevant or redundant features. These features deteriorate the performance of the system. The medical set needs to be preprocessed effectively to eliminate such attributes. Feature selection is an effective way to address the issues of dimensionality and overfitting. Through this process the researchers can identify the significance of attributes in calculating the risk score of the disease. Variable Selection is an equivalent term used for feature selection. These algorithms have many advantages like improvement in accuracy of the prediction model and increase in computational speed. These techniques also provide better visualization of the data . Fig 1 illustrates how feature selection algorithms can be applied to remove the undesirable features from the dataset. Later, Machine learning algorithms are applied to the reduced dataset. Literature supports the existence of 3 varieties of feature selection algorithms namely Wrapper , Filter and Embedded methods.

A. Filter Procedures

In this category, concepts of statistics and correlation criteria (e.g. ANOVA, Chi-square, Pearson’s correlation) are used to determine the usefulness the features for the model in terms of a score. These features are then ranked on the basis of this score. Low rank features are eliminated while the high rank features are selected for building the model. Relief and ReliefF, Chi-square test, t-test ,F-test ,minimum redundancy and maximum relevance(MRMR) etc. fall under the category of filter methods. Relief and MRMR algorithms have been applied in this study.

B. Wrapper procedures

Different subclasses of attributes are created in this technique. Using each subclass of attributes, the prediction models are built and trained. Based on feedforward /backward criteria, the best subclass of the attributes is selected for each learning technique used. Learning algorithms is considered important for these methods of feature selection. Since these methods consider the dependency among the features ,these methods provide better results than the filter methods. However ,these methods are more prone to over fitting. Also, the method needs to be executed again if learning algorithm is changed. Various examples of wrapper feature selection algorithms are genetic algorithm, Randomized Hill-Climbing, Branch-and-Bound Method, Backward elimination Method, Recursive feature elimination method. Genetic algorithm has been employed in this project to identify the important attributes for heart disease dataset.

C. Embedded procedures

These methods comprise of a blend of wrapper and filter algorithms. Depending on the learning techniques, these algorithms work to enhance the performance. The dataset need not be split into training and test dataset and hence, fast results are obtained. These algorithms work better than wrapper methods as these are computationally inexpensive. and are more immune to over-fitting. Since these methods are also dependent on the classifier, the significant features which provide excellent results for one technique ,may not provide good results with a change in classifier. Few examples of embedded feature selection techniques are CART algorithm , least absolute shrinkage and selection operator (LASSO) method, Recursive Feature Elimination Approach, Weighted naïve Bayes etc. belong to the category of embedded methods feature selection. In this study, we have analyzed the LASSO method.

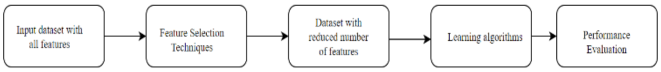
Input dataset with all features
Feature Selection Techniques
Dataset with reduced number of features
Learning algorithms
Performance Evaluation

Figure1: Process for Feature Selection

Table 1 provides the summary of relative advantages and limitations of three categories of feature selection methods

Table 1 Summary of feature selection algorithms

Method	Advantages	Limitation
Filter Method	Computationally competent	Interaction among classifiers is ignored
	Scalable to large datasets	
	Less prone to overfitting	Accuracy is less
	Execution time is less	
Wrapper Method	High predictive accuracy	Computationally expensive
	Detects dependencies among features	Prone to overfitting
	Classifier performance is optimized	Not Suitable for larger datasets
Embedded Method	Accuracy is better than wrapper techniques	Specific to a learning algorithm
	Fast computations	Computationally less efficient than filter methods

III. Methodology

Meaningful insights can be generated from the immense data obtained from the healthcare sector . For this study, we collected heart disease dataset provided by data repository of University of California, Irvine. This is a dataset of 303 records with 14 attributes. The details of this dataset are provided in Table 2. Out of 14 attributes, 13 attributes are the clinical attributes which are linked to diagnosis of heart disease. The output attribute had values from 0 to 4.A value of 0 implies that the person does not have heart disease .A non-zero value is associated with the presence of heart disease in the patient. The higher the value ,the more is the severity of the disease.

Table 2 Details of the heart disease dataset used

S. No	Name of the attribute	Code	Description
1	Age	Age	Age in years

2	Gender	Sex	Female =1
			Male=0

3	Serum Cholesterol	Chol	In mg/dl
4	Fasting blood sugar	Fbs	0=Nondiabetic 1=Diabetic
5	Resting blood pressure	Trestbps	In mm Hg
6	Type of chest pain	Cp	1 =typical angina 2=atypical angina 3 = non-anginal 4= asymptomatic
7	Results of Electrocardiography	Restecg	0 = normal 1 = having ST-T wave defect 2= hypertrophy
8	Maximum value of heart rate achieved	Thalach	Maximum Heart rate
9	Exercise induced angina	Exang	1=Yes 0=No
10	Slope of peak exercise ST segment	Slope	1 = up sloping 2 = flat 3 =down sloping
11	Old peak for ST depression	Oldpeak	ST depression induced by exercise relative to rest
12	Number of major vessels colored by fluoroscopy	ca	Results of fluoroscopy test
13	Thallium scan	Thal	Heart Status 3 =normal 6 =fixed defect 7=reversible defect
14	Diagnosis of heart disease	Num	Presence/absence Of disease 0=absence 1–4= presence of heart disease

The study was carried out with an aim to predict the risk of heart diseases. It is a case of binary classification where we can classify the people in two classes, one with a high risk of the heart disease and other with a low risk of heart disease.

All the computations were performed in Python programming language. The complete study was carried out in 4 steps namely (i) preprocessing the dataset (ii) feature selection using 4 algorithms (iii) Developing the model with 5 classification algorithms (iv) Analyzing the performance of the system.

Fig.2 illustrates the complete methodology used in this project.These steps are elaborated in the following subsections.

A. Preprocessing

Preprocessing of raw data is an important step in developing the prediction models. Optimum results are obtained if the data is processed carefully. It was decided to eliminate all the records with missing values. Standard scaler was applied so that the machine learning algorithms could work effectively. As a result of applying standard scalar, the mean of every feature became zero while the variance was 1.It was observed that the range of attributes were quite different from one another. It is known that classifiers work better if the range of all the features is same. To accomplish this, we employed Min Max Scalar so that all the features are confined in the range 0 to 1.

B. Feature selection

All the attributes in the medical data are not equally important for detecting presence or absence of the disease. Some features hold more importance than the other features. Relatively insignificant features from the data are removed in this step. These algorithms not only help to increase the accuracy of the system but also make the system cost effective as some of the medical tests may be avoided by the patient. There are a variety of feature selection algorithms. In this work, we used applied filter, wrapper and embedded algorithms for feature selection. MRMR and Relief algorithms are two famous filter feature selection algorithms. Genetic algorithm uses wrapper techniques for feature selection while LASSO belongs to the category of embedded techniques.

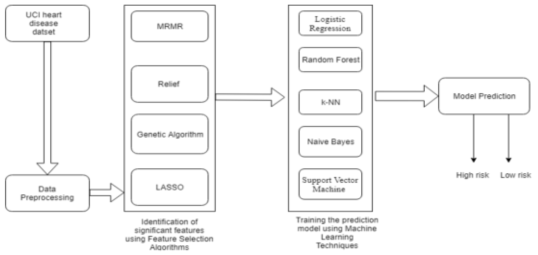

Figure 2: Framework for the prediction system

(1) Relief Algorithm: In this algorithm, the attributes of the dataset are allotted a weightage according to their role in determining the output. The weights are assigned and updated with the passage of time. The more the significance of an attribute, the more is its weight. For each attribute *p*, the following algorithm is iterated through *p* training instances R*(m).* For each *m,* R*(m)* is the output and attribute weight score vector S is restructured.

Need: Training set T consisting of attribute values and the output category value

x ⟵Count of training samples

y ⟵Total count of attributes

p ⟵Count of randomly chosen instances from x used to update S

```
Assume the initial weight score for attribute weights S[A]: =[ 0.0]
    **For** m: = 1 to  p **do**
        Choose  a "target" instance $R_m$ randomly;
        Identify nearest hit and nearest miss .Label them as H and M respectively;
        **For** Y:  1 to y **do**
            S[Y]: S[Y] − diff (Y, $R_m$, H)/p + diff (Y, $R_m$, M)/p;
        **End for**
    **End for**
    Return the weights score matrix S;
```

(2) MRMR: This technique uses optimum attributes which provide least redundancy and are highly relevant for the output attribute. One attribute is checked in a cycle and pairwise redundancy is calculated in terms of mutual information. This algorithm can be represented in the following way:

Input: Count of actual number of attributes in dataset and reduced dataset

Output: selected significant attributes

```
**For**  attribute $A_i$ in original attributes **do**
    Relevance=mutual information ($A_i$, category);
        Redundancy = 0;
        **For** attribute $A_j$ in original attribute **do**
        Redundancy ± mutual information ($A_i$, $A_j$);
        **End For**
        Value[$A_i$] = relevance − redundancy;
**End For**
Sort Values[$A_i$];
Select top order attributes from Values [ $A_i$]
```

(3) Genetic Algorithm: This algorithm has its roots in natural genetics and biological evolution. These techniques work on a population of individuals to produce better and better approximations. Individuals are carefully chosen on the basis of their level of fitness. Recombination is carried out using operators from natural genetics to create new population. The offspring might also undergo mutation. The total number of attributes reflect the number of genes while each individual in the population represents a predictive model. The accuracy of these algorithms is better than the conventional techniques. These algorithms have an added advantage that these algorithms can efficiently process the datasets with large number of attributes. These algorithms provide easy parallelism in computation. However, these algorithms are computationally expensive and are slow.

(4) LASSO: In this algorithm, the absolute value of feature coefficients is updated. The algorithm puts a condition that the sum of absolute values of model parameters cannot be more than a defined upper bound. A regularization strategy is applied in which coefficients of few regression variables are penalized to zero. The attributes whose coefficients become zero are removed from the dataset while the attributes with high values of coefficients are retained in the dataset. This algorithm provides better prediction accuracy and is less prone to overfitting.

C. Training the model

After feature selection, the dataset is divided into two parts called training and test set. The model is built using the training dataset and its evaluation is done using the test dataset. We applied 5 powerful machine learning algorithms namely k-NN, Random Forest, Naïve Bayes, Logistic Regression, Support Vector Machine on the dataset to train the model. The performance of the system was validated by k-fold cross validation technique.

D. Evaluation

It is necessary to check the performance of the model. For this purpose, a we used a number of indices to validate the performance of the system. In this regard, confusion matrix is used to compute performance indices. This is shown in Table 3.

Table 3 Confusion Matrix

		Predicted Risk Score	
		1=High Risk	0= Low risk
Actual Risk Score	1=High Risk	TP	FN
	0=Low Risk	FP	TN

This table is used to compute the following:

TP: True Positive. It is the count of heart patients who were foreseen accurately to have the disease.

TN: True Negative. It is the count of healthy people who were predicted accurately not to have the disease.

FP: False Positive. It refers to those healthy people who were inaccurately predicted to have the illness.

FN: False Negative. It refers to the count of heart patients who have the ailment but are inaccurately predicted to be healthy.

Here, TP and TN, are the cases which were accurately predicted while FN and FP are the inaccurate predictions. It is clear from the table that total number of heart patients is equal to TP+FN while TP+FP represents the number of people who were predicted to have heart disease.

The evaluation parameters used in the study are:

Classification accuracy: It is calculated as the ratio of accurate predictions to the total predictions made. It is calculated as:

(TP+TN)/(TP+FN+FP+TN)

Sensitivity: Sensitivity refers to the fraction of people who were accurately predicted to have heart disease from the total count of people having heart diseases.

Sensitivity =TP/TP+FN

A high value of sensitivity signifies that most patients were accurately predicted to have the heart disease.

Specificity: This is the ratio of accurately predicted cases of healthy people to the actual count of healthy people. This is mathematically calculated as

Specificity =TN/TN+FP

It is desirable to have a high level of accuracy, sensitivity and specificity. The prediction system should have fast processing speed.

IV. Results and Discussions

For this study, we collected Cleveland heart disease dataset from University of California, Irvine data repository. The first step in processing the dataset was to remove the missing values. Out of 303 records, 6 records were removed on account of missing values. All the parameters were standardized and normalized before implementing the classification algorithms. All the computations were carried out in Python programming language. As the first step, we evaluated the performance of all machine learning algorithms with all attributes taken into consideration. Accuracy, specificity and sensitivity of the prediction models were calculated. Later, we applied filter feature selection algorithms (MRMR, Relief), wrapper (genetic algorithms) and embedded feature selection algorithms to identify the significant attributes from the dataset. The insignificant attributes were removed in each case. Prediction models were built on the reduced dataset using 5 commonly used classifiers namely logistic regression, k-NN, random forest, Naïve Byes and Support Vector Machine. Performance evaluation was carried out. K-fold cross evaluation validates the results.

A. Results for the performance of classifiers with all attributes(n=13)

The original dataset has 13 input attributes. In this experiment, all these attributes were utilized and the performance of 5 machine learning algorithms was evaluated. Cross validation was done using k-fold method. The dataset was iteratively split in ratio of 9:1 for training and testing purposes. Performance of 5 classifiers has been presented in Table 4. Accuracy, processing time, sensitivity and specificity were evaluated. It is clear from this table that Support vector machine achieved maximum accuracy of 86.1% and specificity of 88.1%. However, the sensitivity was just 77.8%. The results have been shown graphically in Fig.3. The comparative execution time for all classifiers is represented in Fig.4.

Table 4 Performance metrics of classifiers on all attributes

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	Processing time(s)
Logistic Regression(C=10)	83.8%	84.9	82.8	19.2
k-NN(k=9)	75.8	73.8	72.8	29.4
SVM(kernel =RBF)	86.1	88.1	77.8	15.2
Naïve Bayes	83.1	87.1	78.3	34.1
Random Forest	82.9	70.3	94.0	15.2

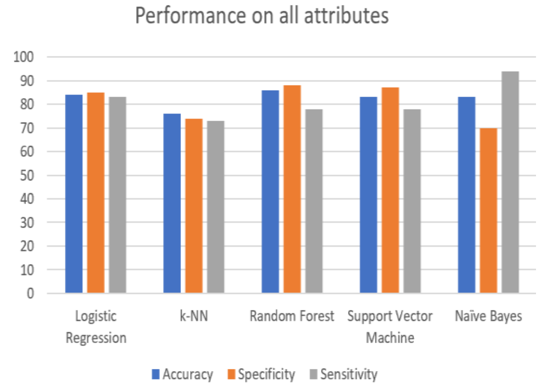

Figure 3: Performance of various classifiers on all attributes of dataset

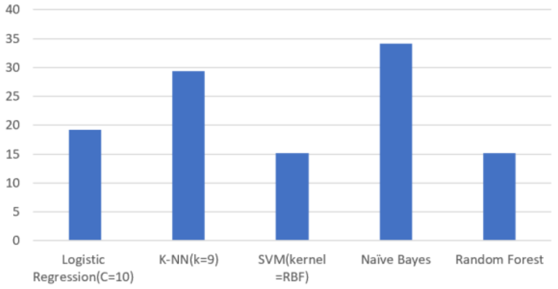

Figure 4: Computation time (s) for classifiers with all attributes

It can be easily inferred that execution time of SVM classifier is 15.2s which is better than the rest classifiers. Considering all these metrics, SVM and logistic regression classifiers ensure promising results as compared to Random Forest, k-NN and Naïve Bayes.

B. Results for performance of classifiers using Relief Algorithm

In this technique, each attribute is associated with a weight. The attributes with higher weights are considered significant while the attributes with lesser weights are considered insignificant and are removed from the dataset. Initially, the models were built using only 3 significant features. Later, the models were built using 6,9 and finally 12 significant features. It was observed that the best results were obtained when 6 attributes were selected. Table 5 depicts the most significant 6 features along with their weights.

Table 5 Significant features obtained for Relief algorithm

Order	Feature	Weight
1	Thallium Scan	0.251
2	Exercise-induced Angina	0.230
3	Chest Pain	0.219
4	Slope of peak exercise ST test	0.129
5	Results of fluoroscopy	0.121
6	Maximum heart rate	0.111

Accuracy, specificity and sensitivity of the prediction models were calculated for the reduced dataset. Further, the execution time was also noted. Table 6 clearly depicts how the performance of the classifiers improves when insignificant features are removed from the dataset.

Table 6 Performance with Relief algorithm

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	Execution time(s)
Logistic Regression(C=100)	88.8	98.1	77.2	16.2
k-NN	80.1	73.2	78.3	24.4
SVM(kernel =RBF,C=100)	87.2	95.0	79.1	14.1
Naïve Bayes	85.3	87.1	78.1	34.1
Random Forest	82.9	93.1	70.2	15.1

Logistic regression (hyperparameter C=100) provided an accuracy of 88.8% when only 6 attributes were considered. It is followed by SVM (kernel=RBF, C=100) which provided an accuracy of 87.2%. The specificity of logistic regression was as high as 98.1% followed by SVM classifier at 95%and Random forest at 93.1%. It can easily be inferred that the computation time for all the classifiers has reduced substantially.

C. Results for performance of classifiers with MRMR Feature Selection

In MRMR feature selection algorithm, selection of the significant features is done according to mutual information between the features. 6 most significant features were selected from the dataset. These significant attributes with their scores are shown in Table 7.

Table 7 Significant features based on MRMR algorithm

Order	Feature	Score
1	Chest Pain	0.590
2	Serum cholesterol	0.575
3	Slope of ST	0.573
4	Result of Fluoroscopy tests	0.541
5	Gender	0.522
6	Thallium Scan	0.487

The attributes listed in Table 7 were retained while the others were ignored. The performance indices of the machine learning techniques were recorded thoroughly. Table 8 highlights the improvement in the performance in terms of accuracy, specificity and sensitivity. Highest level of accuracy was achieved with Naïve Bayes at 84%. Logistic regression (C=100) provided an accuracy of 78% closely followed by SVM (kernel RBF, C=100) with an accuracy of 77%. Specificity of 90% was achieved using Naïve Bayes classifier. Logistic regression and SVM classifiers provided a specificity of 88%. It is observed that the processing speed has increased considerably with MRMR feature selection algorithm. Processing time of Naïve Bayes algorithm was as low as 1.6 seconds. Processing speed of Logistic regression and Random Forest have also improved significantly with MRMR feature selection algorithm. The execution time was 2.2 and 2.0 seconds, respectively.

Table 8 Performance using MRMR algorithm

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	Execution time(s)
Logistic Regression(C=100)	78	88	67	2.2
k-NN(k=7)	62	62	61	10.0
SVM(kernel =RBF,C=100)	77	88	60	60.5
Naïve Bayes	84	90	77	1.6
Random Forest	67	70	62	1.3

D. Results for performance of classifiers with Genetic Algorithm

Application of genetic algorithm to Cleveland heart disease dataset leads to the identification of following significant features: resting blood pressure, Age, fasting blood sugar, type of chest pain, thallium scan, exercise induced angina and old peak. For the development of prediction model, these prominent features were selected while the other attributes were rejected. It is observed that the performance of classifiers improves with the application of genetic algorithms. Table 9 illustrates the effect of genetic algorithm on the performance of classifiers.

Table 9 Performance using Genetic Algorithm

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	Execution time(s)
Logistic Regression(C=10)	85	88	79	15.2
k-NN(k=6)	77	77	71	18.4
SVM(kernel =RBF)	88	89	78	11.2
Naïve Bayes	85	87	85	27.1
Random Forest	82	73	81	11.2

It is clear from these observations that the genetic algorithm helps to make the system more efficient. Accuracy, sensitivity and specificity using SVM classifier were 88%,89% and 78% respectively. The performance of Naïve Bayes and Logistic regression also improved significantly. Accuracy, specificity and sensitivity obtained using Naïve Bayes is 85%,87% and 85% respectively.

E. Results for performance of classifiers with LASSO Feature Selection

Feature selection was carried out to remove irrelevant or redundant features present in the dataset. Feature section algorithm LASSO chooses highly vital features to target and ignores the rest features. The attributes are ranked according to their significance to the output attribute. Fig 5 depicts the weighted scores of the attributes after applying LASSO feature selection.

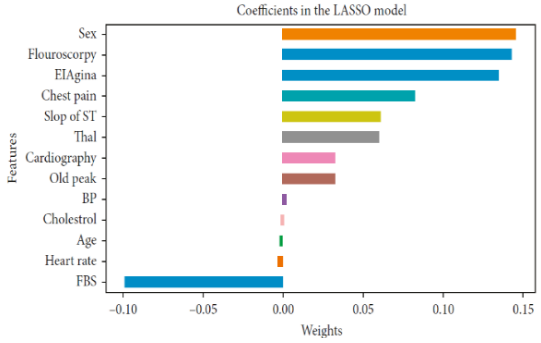

Figure 5: Scores of Attribute Significance with LASSO

The results indicate that the features Gender, Fluoroscopy, EI Angina, Slope of ST and thallium scan are significant in predicting the chance of heart disease.

It is interesting to know here that fasting blood sugar has weak relation to risk score of heart disease in this dataset. The attributes with less importance were removed and only the vital ones were used to make the model. Table 10 represents the performance of the various predictive algorithms on the important attributes. Accuracy, sensitivity and specificity of logistic regression (C=10) were 87%,97% and 76% respectively. The execution was quite fast as the processing time is as low as 0.02 seconds. It is observed that performance of k-NN improved significantly with LASSO algorithm. The best results were obtained with k=3.The sensitivity was found to be 94% while the specificity was 72%. The performance of SVM was found to be optimum for kernel RBF, C=100.The accuracy was observed to be 85%. Sensitivity and specificity were 94% and 74% respectively. The processing time was 0.02 seconds. Fig.6 illustrates graphically the performance metrics of the prediction system designed after implementing LASSO algorithm

Table 10 Performance using LASSO algorithm

Classifier	Accuracy (%)	Specificity (%)	Sensitivity (%)	Execution time(s)
Logistic Regression	87	97	76	0.02
k-NN	83	94	72	0.02
SVM	85	94	74	0.03

Naïve Bayes	83	88	78	6.3
Random Forest	83	92	72	0.02

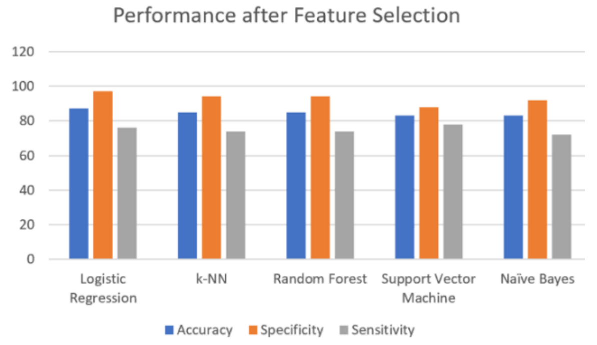

Figure 6: Performance of classifiers with LASSO feature selection

Table 11 Best performance indices and best machine learning algorithms for various feature selection techniques

Best Performance Indices	**Relief Feature selection**	**MRMR feature selection**	**Genetic algorithm**	**LASSO feature selection**
Accuracy	89% Logistic Regression	84% Naïve Bayes	88% SVM	87% Logistic Regression
Sensitivity	98% Logistic regression	90% Naïve Bayes	89% SVM	97% Logistic Regression
Specificity	79% SVM	77% Naïve Bayes	85% Naïve Bayes	78% Naïve Bayes
Execution Time(seconds)	14.1SVM	1.3 Random Forest	11.2 Random Forest	0.03 SVM

The results of this research are summarized in Table 11. An overview of the best performance metrics and the best classifier have been represented here. It can be easily interpreted that the best accuracy on reduced dataset was achieved using Relief feature selection algorithm as compared to other algorithms. Logistic Regression classifier works well with Relief algorithm to achieve the best accuracy of 89% as well as the best sensitivity of 98%. Genetic algorithm of feature selection along with Naïve Bayes classifier achieves the best specificity of 85%. LASSO feature selection algorithm with SVM (kernel=RBF) provides the fastest speed of execution where the processing time is as low as 0.03seconds.

CONCLUSIONS

Heart diseases are responsible for largest number of demises globally. Machine learning based prediction models can forecast the risk of developing heart disease in early stages. The performance of the prediction system can be enhanced by removing redundant and irrelevant attributes from the healthcare data. In this paper 4 feature selection techniques to detect important attributes. These algorithms are MRMR, Relief, genetic algorithm and LASSO. These techniques were used in combination with 5 classifiers namely Logistic regression, k-NN, Naïve Bayes, Random Forest and support vector machine to build a heart disease prediction system. The study has been done on Cleveland heart disease dataset. Results obtained from this research study prove that Relief feature selection with Logistic Regression provides the best results for accuracy and sensitivity of the system. Best specificity is achieved using genetic algorithms with Naïve Bayes classifier. LASSO feature selection with Support Vector machine provides the fastest speed of execution. The results of this study clearly prove that the feature selection algorithms increase the accuracy, sensitivity, specificity and the processing speed of the prediction system significantly. This research can be further extended by implementing other techniques of feature selection to further advance the performance of prediction system.

REFERENCES

[1] Dorairaj Prabhakaran, Kavita Singh,Gregory A. Roth,Amitava Banerjee,Neha J. Pagidipati, Mark D. Huffman."Cardiovascular Diseases in India Compared With the United States".Journal of the American College of Cardiology,Vol.72,No.1,2018

[2] India State-Level Disease Burden Initiative Collaborators.Nations within a nation: variations in epidemiological transition across the states of India, 1990–2016 in the Global Burden of Disease Study. Lancet 2017; 390:2437–60

[3] S. R. Bhagyashree Kiran, Nagaraj ,Martin Prince, Caroline H. D., "Murali Krishna ,"Diagnosis of Dementia by Machine learning methods in Epidemiological studies: a pilot exploratory study from south India"Soc Psychiatry Psychiatr Epidemiol

[4] Nimai Chand Das Adhikari,"Prevention of Heart Problem using Artificial Intelligence" International Journal of Artificial Intelligence and Applications (IJAIA), Vol.9, No.2, March 2018

[5] Maini E., Venkateswarlu B., Gupta A., Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System. In: J. Hemanth et al. (Eds.): ICICI 2018, LNDECT 26, pp. 627–632, 2019. doi:10.1007/978-3-030-03146-6_69

[6] Gupta N, Ahuja N, Malhotra S, Bala A,Kaur G," Intelligent heart disease prediction in cloud environment through ensembling" Expert Systems. 2017;34:e12207.https://doi.org/10.1111/exsy.12207

[7] Kavitha, R., Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1–5

[8] Burak Kolukisa*1, Hilal Hacilar1, Gokhan Goy, Mustafa Kus, Burcu Bakir-Gungor, Atilla Aral, Vehbi Cagri Gungor,"Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease"2018 IEEE International Conference on Big Data (Big Data)

[9] Divya Jain , Vijendra Singh ,"Feature selection and classification systems for chronic disease prediction: A review" Egyptian Informatics Journal 19 (2018) 179–189

[10] Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. Comput Methods Programs Biomed 2016;127:94–104.

[11] Mohammad Shafenoor Amina, Yin Kia Chiama, Kasturi Dewi Varathan,"Identification of significant features and data mining techniques in predicting heart disease" Telematics and Informatics 36 (2019) 82–93

[12] Ali Muhammad Usman, Umi Kalsom Yusof, Syibrah Naim "Cuckoo inspired algorithms for feature selection in heart disease prediction" International Journal of Advances in Intelligent Informatics ISSN 2442-6571 ol. 4, No. 2, July 2018, pp. 95-106

[13] Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II

diabetes databases. In: Systems, man and cybernetics, 2008. SMC 2008. IEEE international conference on. IEEE; 2008. p. 2628–33.

[14] https://archive.ics.uci.edu/ml/datasets/heart+disease G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

AUTHOR PROFILE

Ekta Maini
B. Tech, ME, (PhD)
Her interest areas include Data Mining and Machine Learning. She has participated in many national and international conferences.

Dr. Venkateswarlu. B
He is presently Associate Professor in the Department of Computer Science and Engineering in Dayanand Sagar College of Engineering and has completed his B.E and M.Tech in Computer Science and Technology from Andhra University College of Engineering, Andhra University. He Completed his Ph.D from Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam. He has 14+ years of teaching experience in various Engineering Colleges. His areas of research interest are Data Mining, Soft Computing Techniques, Software Engineering, Image processing and Data Engineering.
.

Dr. Arbind Kumar Gupta
He is presently Professor in the Department of Computer Science and Engineering in Dayanand Sagar College of Engineering and has 19 years of Industry Experience and 10+ years of teaching experience in various Engineering Colleges. His areas of research interest are Rural Healthcare, Synthetic Aperture Radar (SAR), Image Processing, Computer Vision, Pattern Recognition, Imaging Techniques.
.