International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Leveraging Machine Learning Techniques for Improving Heart Disease Prediction Systems Using Feature Selection

Surendra Reddy Vinta[1], Dr. E. Anbalagan[2], Chethan Chandra S Basavaraddi[3], Anni Princy B[4], Rakshit Govind T[5], Neeta Mazumdar[6]

Submitted: 20/04/2023 **Revised**: 17/06/2023 **Accepted**: 27/06/2023

Abstract: Although there have been advancements in the Indian healthcare system over the past few decades, there is still a long way to go before we can claim to have reached world standards. Despite being the second most populated nation, India ranks 143rd out of 195 nations in terms of healthcare infrastructure. Even now, seven decades after India's declaration of independence, the country's healthcare system remains unable to guarantee universal access to care. Access to affordable, high-quality healthcare is still a pipe dream, especially for those who live in rural areas. Not everyone can afford medical services. Private organizations charge a high price for their therapies. No considerable financial assistance is allowed. This study suggests a novel hybrid feature selection method for identifying the most important qualities. Standard feature selection approaches such as Maximum relevance and minimum redundancy (mRMR), Relief, a genetic algorithm, and Least absolute shrinkage and selection operator (LASSO) were compared. A variety of classifiers were used to create a cardiovascular disease prediction system, including logistic regression, Naive Bayes, Random Forest, and support vector machine. This study made use of data from the Cleveland heart disease dataset. According to the results of this research, a Random forest based prediction model trained using characteristics discovered via a new hybrid feature selection may provide the best accuracy and sensitivity. According to the results of the research, applying feature selection algorithms enhances the performance of the prediction system in terms of accuracy, sensitivity, specificity, and throughput.

***Keywords*:** *Relief Algorithm; Genetic algorithm; healthcare infrastructure, Machine-learning method, Chi-square.*

1. Introduction

Coronary heart disease (CHD) is the most prevalent form of heart disease. This illness causes more than 3,50,00 deaths a year. Heart diseases are also responsible for 22% of deaths (in total heart disease deaths) in Asian countries. There are other risk factors for heart disease (blood pressure, diabetes, current smoking, high cholesterol, etc.). Hence, it is difficult to diagnose heart problems [1]. The severity of cardiac disease in humans has been determined using various data mining and neural network techniques. The complexity of CHD disease necessitates careful management because it is a complex condition. Failure to perform early detection may affect the heart or result in unexpected death. For the purpose of identifying various types of metabolic illnesses, the perspective of medical science and information gathering are used [2]. With the aid of past data samples and examples, a system can learn new things without being explicitly programmed thanks to the machine learning technique. Logic is created by machine learning from past data. Many different fields depend heavily on machine learning. The impact on heart disease detection is also demonstrated. AI, which is also regarded as a subset of machine learning, includes deep learning. Numerous more study fields can benefit from the use of deep learning. It is used to forecast heart problems as well [3].

Almost 70% of all fatalities globally are caused by heart disease, more especially cardiovascular disease (CVDs), which is a primary cause of morbidity and mortality. A study indicates that more than 43% of deaths are caused by CVD. In high-income countries, unhealthy eating, tobacco use, too much sugar, and being overweight or

1Associate Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India. Email: vsurendra.cse@gmail.com Orchid Id: https://orcid.org/0000-0002-5882-733X
2 Professor, Department of big data and network security, SSE, SIMATS School of Engineering, Chennai, Tamil Nadu, India. Email Id: eanbalagan77@gmail.com
3Associate Professor, Dept. of Computer Science & Engineering, Kalpataru Institute of Technology, Tiptur, Karnataka, India. Email Id: raddi04@yahoo.com Orchid Id: https://orcid.org/0000-0002-3133-7234
4Professor, Computer and Communication Engineering, Panimalar Engineering College, Chennai, Tamilnadu, India. Email : ccehod@panimalar.ac.in ORCID ID : https:// orcid.org/0000-0002-1464-1402
5Student, ECE, BMS Institute of Technology and Management Email Id: rakshitgovind982@gmail.com Orchid Id: https://orcid.org/0009-0007-6670-3234
6Associate Professor, Department of Mathematics, Government College of Arts, Science and Commerce Quepem Goa, India. Email Id: neetamazumdar@gmail.com Orchid Id: https://orcid.org/0000-0002-8971-1065

having extra body fat are common risk factors for heart disease [4]. Yet, the prevalence of chronic diseases is also rising in low- and middle-income nations. Furthermore, many low- and middle-income countries find it difficult and frequently prohibitively expensive to get diagnostic tools like electrocardiograms and CT scans, which are crucial for finding coronary heart disease. The physical and financial toll that heart disease has on people and businesses must thus be reduced by early detection. A WHO assessment predicts a significant rise in CVD-related mortality overall, primarily from heart disease and stroke. Hence, in order to save lives and lessen the financial burden on society, it is essential to apply data mining and machine learning approaches to anticipate the possibility of getting heart disease [5].

Investigating hidden patterns in data sets in the healthcare industry is an excellent application for machine learning. Machine learning and data mining techniques have made it possible to extract pertinent patterns and identify correlations and relationships between different clinical variables. With the aid of these cutting-edge and complex algorithms, large amounts of data may be simply analysed [6]. These technical developments can be applied to the healthcare industry to collect, arrange, and systematically analyse clinical data. A platform for a better understanding of the mechanisms in all facets of patient care will be created by the combination of these technical breakthroughs, which will also result in enhanced early diagnosis, superior medicine, and successful treatment [7]. In brief, the pertinent data will be pulled from the medical databases to help with early disease detection and ultimately disease prevention. Researchers are drawn to creating diagnostic tools through data mining and machine learning techniques. In the recent past, wealthy nations have proposed machine learning-based prediction systems for early diagnosis of cancer, CVDs, Parkinson's, dementia, diabetes, etc [8].

Using data mining techniques, the medical industry generates a considerable amount of data every day, and we can uncover hidden patterns that can be applied to clinical diagnosis. As a result, data mining is crucial in the medical industry, as demonstrated by research done over the previous few decades [9]. When forecasting cardiac disease, a number of variables need to be taken into account, including diabetes, high blood pressure, high cholesterol, and abnormal pulse rate. The outcomes in forecasting cardiac disease are frequently impacted by the incompleteness of the medical data supplied. In the medical industry, machine learning is extremely important. We are able to identify, detect, and predict a variety of diseases using machine learning [10]. Using data mining and machine learning techniques to forecast the chance of contracting specific diseases has recently attracted increasing interest. Data mining techniques are used in the already published study to forecast the disease. Although some studies have attempted to forecast the likelihood that the disease would proceed in the future, they have not yet produced reliable results. This paper's major objective is to accurately forecast the likelihood of cardiac disease in the human body [11].

What follows is the outline for the rest of the paper. The related work is briefly described in part 2, and the methodology and the theoretical foundations of the methods used are described in section 3. The simulation results and analysis are presented in section 4. For the chapter's final section, "key findings" we summarize the most important results.

2. Past Related Work

Regression and classification are the two key tasks where machine learning techniques find their use. The classification of diseases using machine learning is a typical example. Here, a number of elements are taken into account, and the outcome is based on those aspects [12]. Features/attributes are the terms used to describe these input variables. High-dimensional data refers to a dataset with several input features, typically on the order of 100 or more. The "Curse of Dimensionality" is a group of issues that are typically linked to high dimensional data [13]. Finding patterns in the data is challenging when there are many input attributes (high dimensionality). Data Sparsity is the name for this dimensionality curse. Sparse data during ML model training results in overfitting and high variance. The amount of training data must rise exponentially in order to generalize the trend when features and characteristics are added [14].

Many studies on feature selection methods have been published in the literature. The importance of feature selection in the dimensionality reduction of data was summed up in a study conducted by a few scholars [15]. The numerous feature selection approaches are also summarized in this paper. A different group of academics presented a mutual information-based strategy with the aim of maximizing target relevance and reducing attribute redundancy [16]. This research calculates the mutual information between each feature in a dataset. Several academics conducted a study that explored the difficulties associated with the high dimensionality of the data. This study also looked at the effect of feature choice. The paper provided a thorough description of the theoretical underpinnings, practical difficulties, and applications of feature selection [17].

Many researchers studied and contrasted various dimensionality reduction methods applied to microarray datasets. The authors conducted a thorough analysis of

feature selection and feature extraction and explained techniques to increase classification accuracy and reduce computing complexity. Another group of academics presented an efficient statistical methodology for the feature selection and classification of gene expression data [18]. The goal of the study was to identify the most relevant genes responsible for the existence of illnesses. The study's experimental findings, which were remarkably accurate, show that feature selection enhances system performance [19]. A team of experts conducted research to create a precise diagnostic system for knee joint diseases. The best performance, or accuracy of 94.31% in terms of accuracy, precision, and recall, was obtained using the proposed technique. The early detection of knee-joint diseases can be accomplished with the help of this investigation. Early diagnosis makes it possible to provide patients with treatment at an early stage [20].

Several writers have performed fundamental groupings of feature selection and various gene selection methodologies. Different methods of feature selection were divided into three categories: supervised, unsupervised, and semi-supervised. This work also addressed a number of difficulties and impediments to learning from gene expression data. 1) How to cope with noisy and incorrectly classified data, as well as 2) how to deal with extremely imbalanced data, are some of the key questions. 3) How to choose the important attributes; 4) How to determine the relevance or redundancy of the attributes [21]. Another team of researchers used a novel feature selection technique that combined SVM ranking with a backward search strategy to find the most useful significant features for the type 2 diabetes dataset. This method significantly improved the Naive Bayes classifier's accuracy. The findings of this study are very helpful to doctors in the early detection of Type 2 diabetes [22].

A few researchers have suggested the quick and effective dimensionality reduction technique known as Modified FAST. The correct discovery of a perfect threshold value with symmetric uncertainty (SU) was made. The application of symmetric uncertainty led to the construction of the least spanning tree (SU). The proposed method's outcomes were contrasted with those of other algorithms like Relief, FCBF, and CFS. According to classification accuracy and the proportion of qualities that were chosen, Modified FAST outperformed the others [23]. A popular feature selection method is known as Mutual Information-based Feature Selection or MIFS. A team of academics devised this approach. The idea of mutual information is used to conduct a "greedy" search for qualities. The process of feature identification is designed to gather the most information from multiple sources. MIFS-U, a modified version of MIFS, was developed to fully utilize the mutual information between data elements [24]. When there is a uniform distribution of information, its performance resembles an "ideal greedy selection algorithm." However, it has been noted that MIFS performance suffers when data element information distributions deviate from a uniform distribution. The Mutual Information-based Constructive Criterion (MICC) technique was created to address this problem. This approach, a greedy filter feature selection, takes into account the attributes' non-redundancy as well as their relevance to the output class. Compared to its predecessors, MICC performs better.

3. Purpose of the work

1) To suggest new hybrid feature selection and ensembling techniques for improved prediction system performance.
2) To determine the main challenges to seamless integration of machine learning in the current Indian healthcare industry.

4. The Proposed Work:

The purpose of this research was to evaluate and suggest an improvement upon existing feature selection methods by utilising a novel hybrid feature selection algorithm.

4.1. Pre-processing

The raw data must be pre-processed before the prediction models can be created. If the data is processed with care, the best outcomes can be achieved. The decision was made to remove all records that were incomplete. The machine learning techniques were optimized by applying a standard scaler. When standard scalar was applied, all features' means became zero and their variances became one.

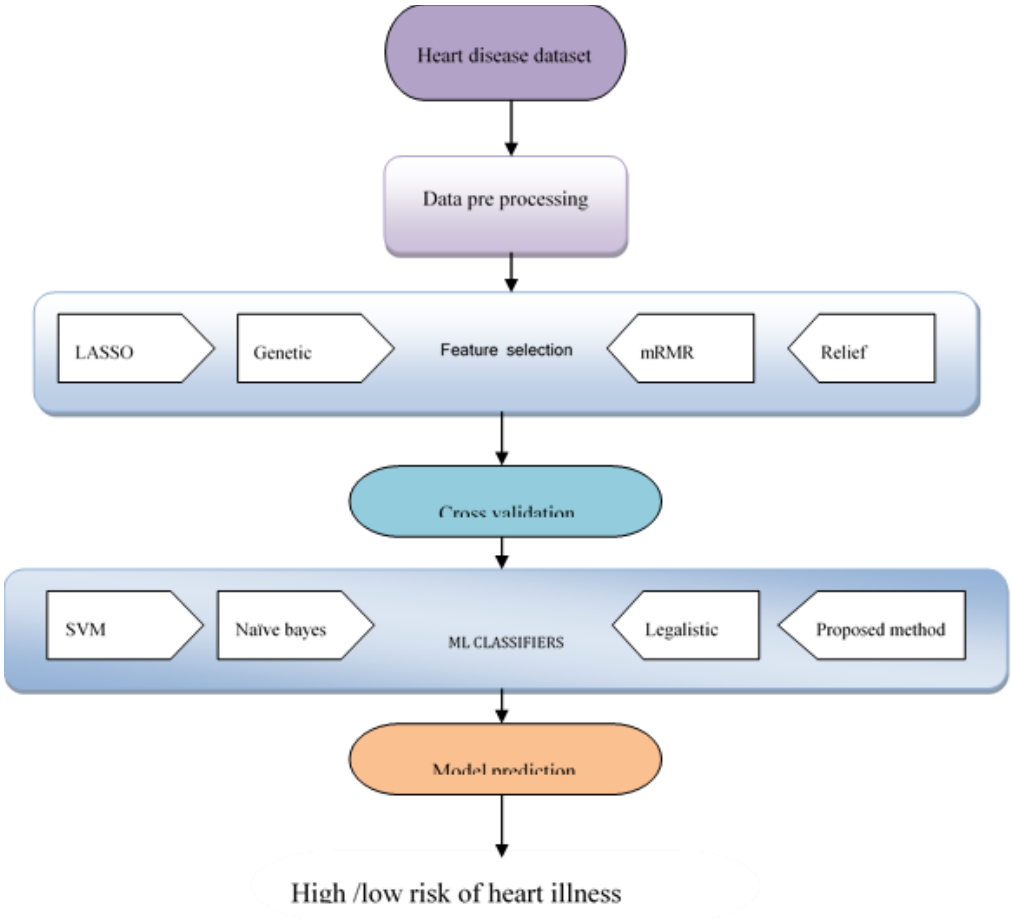

Fig 1: The Proposed System block Diagram Novel Hybrid Feature Selection

4.2. Feature Selection:

Not all characteristics in the medical data are equally useful in making a diagnosis of disease. There are some characteristics that are more vital than others. This process gets rid of the data's less important details. These algorithms not only improve the system's accuracy, but also reduce costs by allowing patients to forego unnecessary medical testing.

Different feature selection algorithms exist. Feature selection was accomplished through the use of an applied filter, a wrapper, and an embedded algorithm in this study. The mRMR method, which minimises duplication and maximises relevance, and the Relief algorithm are two well-known examples of feature selection filters.

To determine which features are most useful when developing a system to predict the likelihood of heart disease, a novel hybrid approach has been presented. Chi-square analysis, information gain, and principal component analysis are only few of the statistical methods used to identify critical characteristics.

4.2.1. Chi-square test: The chi-square statistic, denoted as, is used to examine whether or not two events are related.

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (1)$$

Where, O is observed value and E is expected value.

4.2.2. Information Gain: They are frequently employed as branching conditions in decision trees. Entropy can be evaluated as:

$$Entropy = -\sum P \; log_2 P \quad (2)$$

If p is the proportion of instances in a class, then.

Entropy decreases as a result of information gain. The quality that decreases entropy and hence promotes information gain is prized. Therefore, knowledge gain is a powerful method for selecting features. The entropy before and after a transformation is applied can be used to calculate the information gain. It can be mathematically computed as the change in entropy H from one state to another.

$$IG\ (X,n) = H(X) - H(X/N) \quad (3)$$

4.2.3. PCA: Principal component analysis (PCA) is a popular method for doing so. The goal is to find the major components that account for the most of the variation in the data. A New Hybrid Feature Selection Approach:

1: Normalization: Chimax represents the maximum chi square value. To standardise 2 test values, we divide the remaining values by Chimax.

2. Put the results of the statistical tests in ascending order from lowest to highest. These numbers are listed from highest to lowest.

3. Using the merge sort method, the values are combined into a single list.

4. Pick the Important Characteristics the 'n' most important characteristics can be picked from the combined list.

4.3. Model Training: After the features have been selected, the dataset is split into a training set and a test set. The model is developed with the help of the training dataset, and its performance is measured with the help of the test dataset. To train the model, we used the dataset in conjunction with five sophisticated machine learning algorithms: k-NN, Random Forest, Naive Bayes, Logistic Regression, and Support Vector Machine. The k-fold cross validation method was used to verify the system's accuracy.

5. Result and Analysis:

Validating the model's efficacy is essential. Here, we employed several indicators to verify the system's efficacy.

5.1. Accuracy:

It is a typical metric for classifying test results numerically. Increased precision indicates a more efficient system.

$$\text{Accuracy} = \frac{TN+TP}{Total\ data\ Sample}\ X100$$

(4)

5.2. Recall:

Specificity was defined as the absence of incorrect data classification. True Negative Rate is another name for it (TNR). Figure IV displays the recall of the current method in comparison to commonly utilized methods.

$$\text{Specificity} = \frac{TN}{TN+FP}\ X100$$

(5)

5.3. Sensitivity:

The accuracy with which the model places the test data into one of its classes constitutes the present method's sensitivity. How many true positives were successfully detected was the question it addressed. True Positive Rate is another name for it.

$$Senstivity = \frac{TP}{TP+FN}\ X100$$

(6)

The cases labelled TP and TN were correctly predicted, while those labelled FN and FP were not. The table clearly shows that TP+FN equals the overall number of heart patients, whereas TP+FP stands for the expected number of people with heart disease.

Python was used to perform all of the calculations. First, we compared the efficacy of several machine learning techniques, taking into account all relevant parameters. The prediction models' accuracy, specificity, and sensitivity were computed. Subsequently, we used filter feature selection methods (minimal redundancy maximum relevance, Relief), wrapper (genetic algorithms), and integrated feature selection algorithms to isolate the most important characteristics of the dataset. In each example, the irrelevant details were eliminated.

Table 1: Evaluation of Accuracy (%) for different classifiers.

Classifiers	**Accuracy (%)**				
	Relief algorithm	mRMR Algorithm	Genetic Algorithm	LASSO algorithm	**Proposed Algorithm**
Logistic Regression	**88.2**	**78.5**	**85.3**	**87.2**	**87.8**
SVM	**87.6**	**77.9**	**88.1**	**85.5**	**86.7**
Naïve Bayes	**85.3**	**84.6**	**85.7**	**83.7**	**84.6**
Random Forest	**82.9**	**68.4**	**82.9**	**83.2**	**89.8**

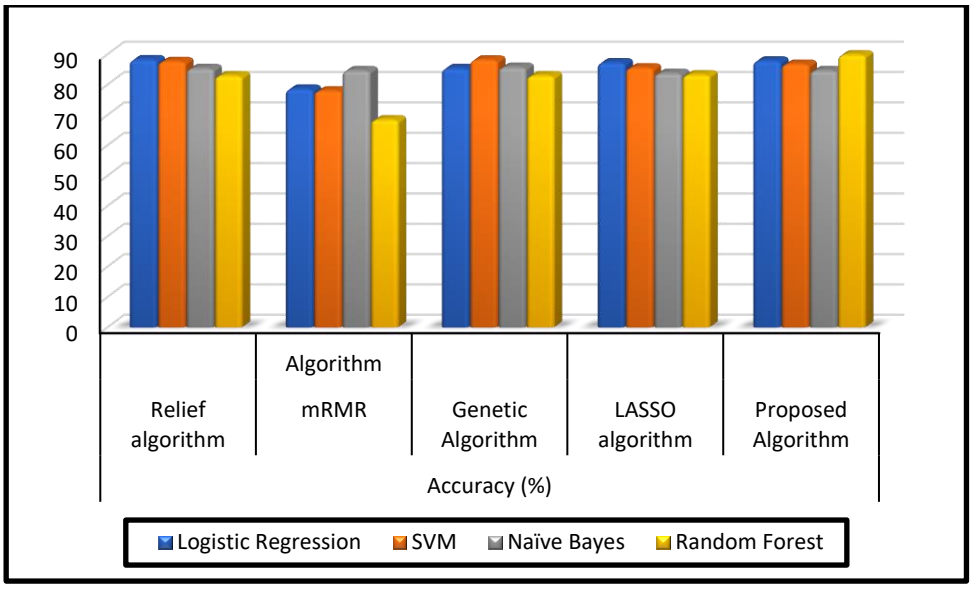

Fig 2: Evaluation of Accuracy (%) for different classifiers.

Table 2: Evaluation of Specificity (%) for different classifiers.

Classifiers	Specificity **(%)**				
	Relief algorithm	mRMR Algorithm	Genetic Algorithm	LASSO algorithm	**Proposed Algorithm**
Logistic Regression	**98.2**	**88.4**	**76.8**	**97.3**	**96.2**
SVM	**95.8**	**88.7**	**75.4**	**94.5**	**94.8**
Naïve Bayes	**88.2**	**90.4**	**73.8**	**88.7**	**88.1**
Random Forest	**93.6**	**90.7**	**71.9**	**92.7**	**97.4**

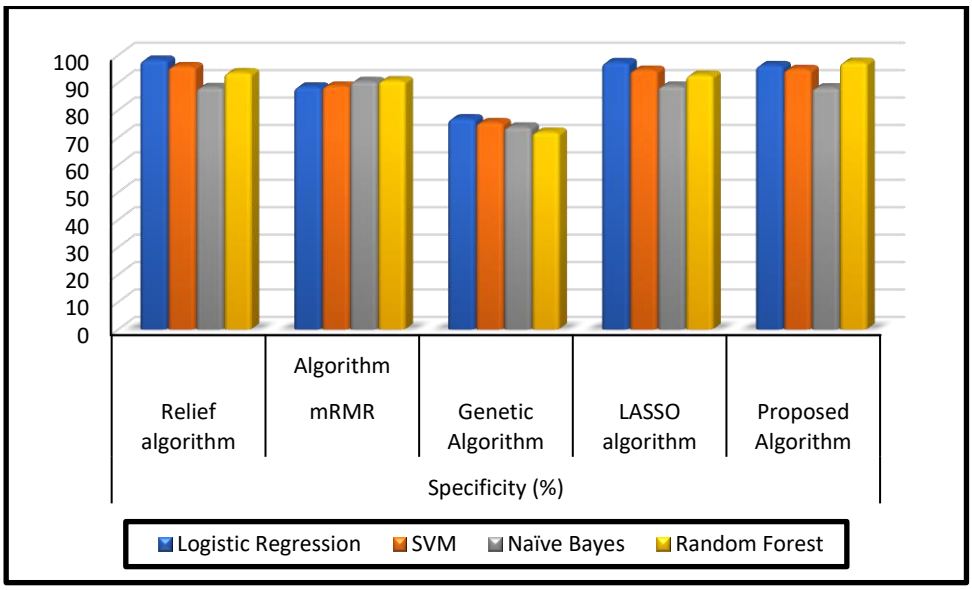

Fig 3: Evaluation of Specificity (%) for different classifiers.

Table 3: Evaluation of Sensitivity (%) for different classifiers.

Classifiers	Sensitivity **(%)**				
	Relief algorithm	mRMR Algorithm	Genetic Algorithm	LASSO algorithm	**Proposed Algorithm**
Logistic Regression	**77.6**	**68.4**	**79.2**	**76.7**	**77.8**
SVM	**79.3**	**67.2**	**78.6**	**74.8**	**73.5**
Naïve Bayes	**78.4**	**78.4**	**85.7**	**78.6**	**74.8**
Random Forest	**71.8**	**62.9**	**81.9**	**72.7**	**78.2**

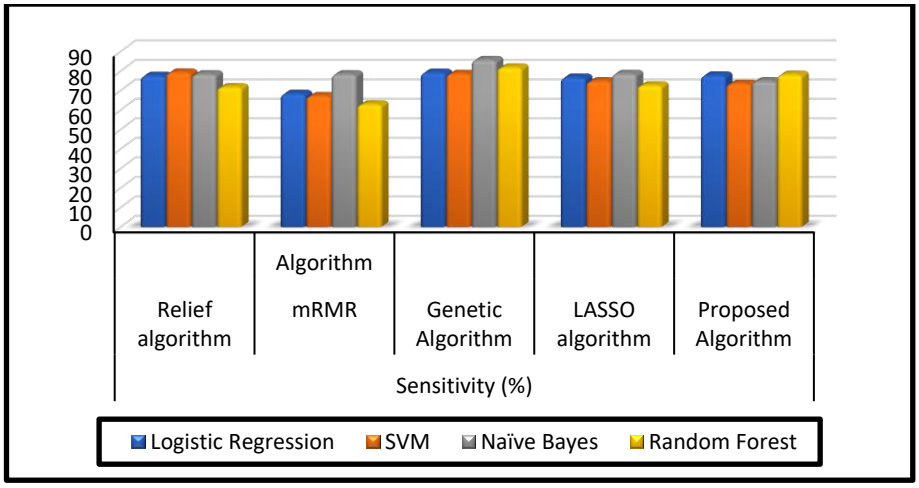

Fig 4: Evaluation of Sensitivity (%) for different classifiers.

Here, we have a summary of the top classifiers and performance indicators. A novel hybrid feature selection technique outperformed competing methods in terms of accuracy on the condensed dataset. When the model was trained on the critical attributes chosen with the help of the innovative hybrid approach, the Random Forest technique was able to attain an accuracy of 89.8 percent. The Relief method with the Logistic Regression classifier produce the highest accuracy (88.2%) and sensitivity (79.3%) possible.

5. Conclusion:

The leading cause of death around the world is cardiovascular disease. Prediction models powered by machine learning can provide early warning of a person's risk of acquiring heart disease. Removing superfluous information from healthcare data can improve the prediction system's efficiency. In order to determine which characteristics are crucial, this research proposes a novel hybrid feature selection technique. Maximum relevance and minimum redundancy (mRMR), Relief, a genetic algorithm, and Least absolute shrinkage and selection operator (LASSO) were compared as standard feature selection methods. Logistic regression, Naive Bayes, Random Forest, and support vector machine were some of the classifiers employed in the development of a cardiovascular disease prediction system. The Cleveland heart disease dataset was used for this analysis. This study's findings show that the best accuracy and sensitivity may be achieved with a Random forest based prediction model trained on attributes found via a new hybrid feature selection. The study's findings show that using feature selection algorithms improves the prediction system's accuracy, sensitivity, specificity, and processing speed.

References

[1] The Lancet, "Health in India, 2017", The Lancet, vol. 389, no. 10065, p. 127, 2017. Available: 10.1016/s0140-6736(17)30075-2

[2] Who.int, 2021. [Online]. Available: https://www.who.int/hrh/resources/16058health_workforce_India.pdf. [Accessed: 20- March- 2021].

[3] Niti.gov.in, 2021. [Online]. Available: https://www.niti.gov.in/sites/default/files/2020-12/PHS_13_dec_web.pdf. [Accessed: 28- January- 2021].

[4] Kasthuri A. “Challenges to Healthcare in India - The Five A's. Indian journal of community medicine “: official publication of Indian Association of

Preventive & Social Medicine, 43(3), 141–143. https://doi.org/10.4103/ijcm.IJCM_194_18

[5] V. Bajpai, "The Challenges Confronting Public Hospitals in India, Their Origins, and Possible Solutions", Advances in Public Health, vol. 2014, pp. 1-27, 2014. Available: 10.1155/2014/898502

[6] Agarwal R, Mittal M. Inventory classification using multi-level association rule mining. Int J Dec Supp Syst Technol. (IJDSST), 2019;11(2):1–12.

[7] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of 20th international conference very large data bases, VLDB. Vol. 1215, pp. 487–499; 1994.

[8] Akbaş KE, Kivrak M, Arslan AK, Çolak C. Assessment of association rules based on certainty factor: an application on heart data set, in 2019 International artificial intelligence and data processing symposium (IDAP) (pp. 1–5). IEEE; 2019.

[9] Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: a survey. Artif Intell Rev. 2017;47(3):313–40.

[10] Alwidian J, Hammo BH, Obeid N. WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl Soft Comput. 2018;62:536–49.

[11] American Heart Association. Heart disease and stroke statistics 2017 at-a-glance. Geraadpleegd van: https://healthmetrics.heart.org/wp-content/uploads/2017/06/Heart-Disease-and-Stroke-Statistics-2017-ucm_491265.pdf.

[12] Amin MS. Identifying significant features and data mining techniques in predicting cardiovascular disease; 2018.

[13] Amin MS, Chiam YK, Varathan KD Identification of significant features and data mining techniques in predicting heart disease. Telem Inform. 2019;36;82–93.

[14] Repository.upenn.edu, 2021. [Online]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1176&context=hcmg_papers.

[15] P. Arokiasamy, "India's escalating burden of non-communicable diseases", The Lancet Global Health, vol. 6, no. 12, pp. e1262-e1263, 2018. Available: 10.1016/s2214- 109x(18)304480

[16] "Lifestyle diseases in India", Pib.gov.in, 2021. [Online]. Available: https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1540840. [Accessed: 28- May- 2021]. [10]"Non-communicable diseases", WHO. int, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[17] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 619–623). IEEE.

[18] Cengiz AB, Birant KU, Birant D. Analysis of pre-weighted and post-weighted association rule mining, in 2019 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1–5). IEEE.

[19] Chauhan A, Jain A, Sharma P, Deep V. Heart disease prediction using evolutionary rule learning, in 2018 4th International conference on computational intelligence & communication technology (CICT) (pp. 1–4). IEEE; 2018.

[20] Dey L, Mukhopadhyay A. Biclustering-based association rule mining approach for predicting cancer-associated protein interactions. IET Syst Biol. 2019;13(5):234–42.

[21] "India: Health of the Nation's States", Institute for Health Metrics and Evaluation, 2021. [Online]. Available: http://www.healthdata.org/policy-report/India-health- nation%E2%80%99s-states.

[22] "National Program for Prevention and Control of Cancer, Diabetes, CVD and Stroke(NPCDCS) | National Health Portal Of India", Nhp.gov.in, 2021. [Online]. Available: https://www.nhp.gov.in/national-programme-for-prevention-and-control- of-c_pg. [Accessed: 28- May- 2021].

[23] "Ayushman Bharat - National Health Protection Mission | National Portal of India", India.gov.in, 2021. [Online]. Available: https://www.india.gov.in/spotlight/ayushman-bharat-national-health-protection-mission.

[24] Fitriyani NL, Syafrudin M, Alfian G, Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access. 2020;8:133034–50.

[25] Kevin Harris, Lee Green, Juan Garcia, Juan Castro, Juan González. Intelligent Personal Assistants in Education: Applications and Challenges. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/185

[26] Nayak, R. ., & Samanta, S. . (2023). Prediction of Factors Influencing Social Performance of Indian MFIs using Machine Learning Approach. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 77–87. https://doi.org/10.17762/ijritcc.v11i1.6053