Webology (ISSN: 1735-188X)
Volume 18, Number 4, 2021

# Predicting Heart Disease Using Feature Selection Techniques Based On Data Driven Approach

**S.USHA[1] , Dr.S.KANCHANA[2]**

[1]Research Scholar, Department of Computer Science SRM Institute of Science & Technology Kattankulathur

[2]Assistant Professor, Department of Computer Science SRM Institute of Science & Technology Kattankulathur

---

**Abstract**
Machine learning techniques, a type of artificial intelligence, are being used in the health field to assist researchers in recognizing pathology before it becomes a major problem. Because healthcare is such an important aspect of a country's economy, researchers are exploring the level of uncertainty that arises when using machine learning algorithms for ways to anticipate the disease. The most significant concept in health data analysis is the prediction of cardiac disease from clinical data. The prediction helps physicians to take exact decisions regarding patients' health. The proposed model used Data collection, Data pre-processing, and Data Transformation methods to train the model. This model exploited feature selection methods: filter and wrapper with classification techniques to enhance the prediction of cardiac disease classification. The classification techniques, namely: Decision Tree, Logistic Regression, Random Forest, and Ada Boost are pragmatic to evaluate performance metrics. The performance metrics include Accuracy, F1-score, Precision, Sensitivity; Specificity reveals an improvement in the outcomes of the prediction.

**Keywords:** Decision Tree, Logistic Regression, Random Forest, and Ada Boost, Filter, Wrapper

## 1. Introduction

A scarcity of clinical specialists, a growth in the number of chronic diseases, and rising healthcare costs are all barriers in today's environment. Heart disease remains the leading cause of premature mortality. Heart diseases occur when enough blood does not reach the body's needs throughout the pumping process [Zheng, Y.(2018)]. Because of multiple contributing danger issues such as diabetes, high blood pressure, high cholesterol, incorrect pulse rate, and many other conditions, it is difficult to detect heart disease. People's health,

particularly their hearts, suffers as a result of busy lifestyles and junk food consumption. An accurate decision support system can play a key role in the early-stage identification of heart problems in developing countries when heart cardiologists are still not available in remote, semi-urban, and rural locations [(Rani et al., 2021)].

Machine Learning and AI techniques have become more popular in healthcare in recent years, particularly for disease diagnosis and risk prediction. This cutting-edge technology will boost processing efficiency while existing technology-based resources will provide a smart way to track and collect data on patient health. To develop accurate forecasts regarding a specific condition, data from diverse sources must be combined with machine learning and artificial intelligence algorithms [Rath (2019)]. Years of medical data collection have given clinicians a new way to diagnose patients. [Escamila et al. (2020)].

A variety of methods are used to categorize the harshness of the disease, including the K-Nearest Neighbor Algorithm, Decision Trees, Genetic Algorithm, and Naive Bayes [Durairaj M & Revathi V, (2015) and Gavhane et al., (2018)]. The severity of cardiac disease in humans has been determined using a variety of ways. Two feature selection strategies are proposed to aid an automatic coding algorithm that has been trained on different data sources. These techniques are used to decrease redundancy between features collected from several sources of data and build a more dense representation of the data with little losing quality.

## 2. Related Works

[Jabbar et al (2016)] proposed work employed RF to predict cardiac illness. The CHI approach was utilized to choose to take the related features. When compared to decision trees, the proposed research suggests that random forests yield more accurate results. The proposed work was built utilizing neural networks by [Kim JK, Kang S (2017)]. Sensitivity analysis is indeed one of the evaluation metrics for prediction. The importance of features with such a high degree of sensitivity was considered. After selecting the relevant characteristic, correlated features were used to examine changes insensitivity. The sensitivity of each feature is determined by it. This [Amin Ul Haq (2018)] employed seven classification algorithms to predict cardiac disease in people. This study used Relief, MRMR, and LAS, and Selection Operator feature selection methods to choose the appropriate feature. In addition to the seven performance metrics this study employed, the ROC and AUC will help clinicians diagnose heart patients more efficiently. To select an appropriate feature, [Rani et al., (2021)] used a Genetic Algorithm (GA) and recursive feature elimination. The proposed study used standard and SMOTE to preprocess the data and performed support vector machines, naive Bayes, logistic regression, random forest, and an Ada Boost classifier to aid in the earlier prediction of heart disease hung on the patient's medical features. The system's simulation environment was built in Python, and it was discovered that random forest achieved a maximum accuracy of 86.6 percent. [Ali et al., (2019)] used the chi-square statistical approach to pick significant features. Particular features that were selected were fed into a deep neural network, which was then trained to do classification. A rigorous grid search method would be used to improve network configuration. [Paul et al. (2016)] used a fuzzy decision support system (FDSS) that includes rules derived from the genetic algorithm with perhaps even weighted fuzzy derivatives (GA). They were able to recover eight useful features with an accuracy of 80%. Multiple heart disease datasets were employed in this

study [Bashir et al., (2019)] for experimentation analysis and to increase accuracy performance. Feature selection algorithms such as Decision Tree, Logistic Regression SVM, Nave Bayes, and Random Forest are used with the Rapid miner, and accuracy is enhanced. [Liu et al. (2017)] offered a study that used relief and rough set approaches. The proposed system consists of two subsystems: the RFRS feature system and ensemble classifier classifications. The first system has three stages: data extraction using the ReliefF method, feature reduction using our heuristic Rough Set reduction technique, and feature reduction using our heuristic Rough Set reduction technique. In the second system, which is based on the C4.5 classifier, an ensemble classifier is suggested. The proposed technique had a classification accuracy of 92.32 percent. On the Cleveland heart disease dataset, [Singh et al.(2017)] used an RF classifier that can handle large amounts of data with missing values. This classifier generates a large number of decision trees that are selected through voting. The chosen branch is used to improve precision. Due to the obvious non-linear dataset, this study was able to reach an accuracy of 85.81 percent.

## 3. Classification Models

Decision Tree, Logistic Regression, Random Forest, and Ada Boost are four machine learning methods that were used in this study.

.

### 3.1. Decision Tree

The decision tree is one of the most well-known machine learning approaches (DT). It is often used in classification techniques. A decision tree represents the decision logic for classifying data objects. It is depicted as a tree, with internal nodes representing attributes, branches representing decision rules, and leaf nodes representing ultimate outcomes. The root node is the initial or top-most node in a DT tree, and it usually has multiple levels. Input variable or attribute testing is reflected in all internal nodes (those with at least one child). Based on the experiment result, the classification algorithm branches in the direction of the appropriate child node, and the procedure of testing and branching continues till it reaches the leaf node. [Quinlan JR (1986)]. DTs are easy to read and understand, and they are employed in a variety of medical diagnostic procedures. When traversing the tree for classification, the results of all tests at each node along the path will provide enough information to make an educated guess on the class of the sample.

### 3.2.Logistic Regression

LR is a simple supervised learning method. It is mostly used to solve problems involving binary categorization. It's a type of ordinary regression that can only model a binary variable, such as whether or not an event occurs. Logistic Regression may help you figure out if a new instance belongs to a given class. The result will be between 0 and 1 because it is a probability. When employing LR as a binary classifier, a threshold must be chosen to discriminate between two categories. Multi-valued variables can be modeled using the LR method. Multinomial logistic regression is a larger variant of LR. [Uddin et.al., (2019)]

### 3.3. Random Forest

Random Forest is a tree-based method for classification as well as regression analysis. RF is an ensemble strategy in which the dataset is divided into small subsets and the decision tree algorithm is applied to each of them. Sampled-with-replacement is used to sample the subgroups. Voting determines the relevance of features. The bagging technique or bootstrapping is a technique that minimizes variance without influencing the bias of the entire ensemble. Out-of-bag scores are used to authenticate the results. Overfitting and large variance in outcomes are reduced with this strategy. [Breiman L., (2001)]

### 3.4. Ada Boost

Ada Boost is a classification algorithm. It creates numerous weak learners and then combines them to create a powerful classifier. This process illustrates that an error made by one learner has an impact on other learners' errors. This process of building weak learners and adjusting sample weights continues until the weak learners' conditions are met. The voting power of the classifier and the classifier's decision criteria are used to calculate the function. The maximum vote received by the label when the predictive function is applied is used to determine the test sample's final prediction. In Ada Boost, the classifier's correct predictions have greater voting power than the classifier's major misclassifications.

## 4. Methodology

In this study's technique, the dataset for classification is acquired initially. One of the most important aspects of Machine Learning is classification. Assigning a document around one or more groups or categories is known as classification. Therefore, the suggested work uses the feature selection method chi-square and recursive feature reduction to show the outcomes of all four application algorithms: Decision Tree, Logistic Regression, Random Forest, and AdaBoost. This research's work is depicted in fig 1.

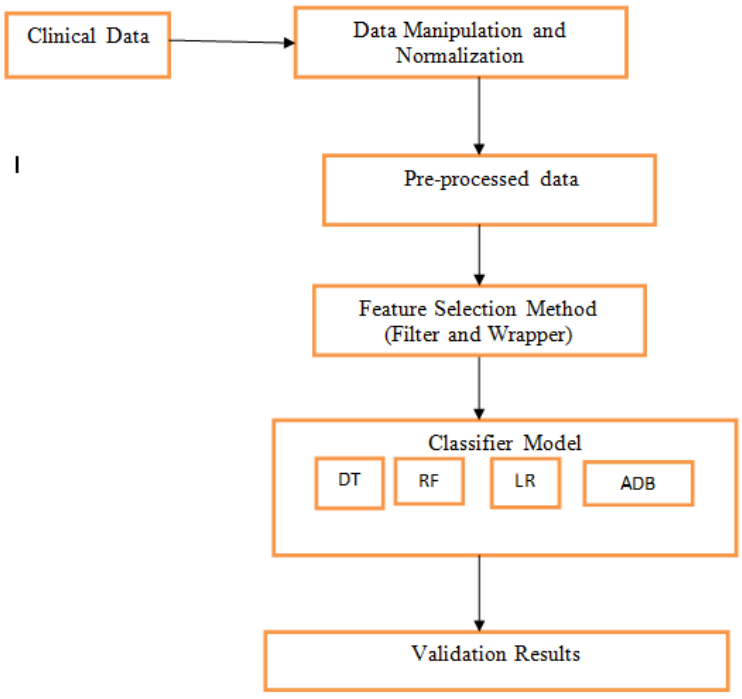

**Figure 1 Proposed Work**

### 4.1. Description of Dataset

The dataset was collected online in the Kaggle repository. It is used to analyze the models chosen for this work. There are 70000 patient records in the collection, each with 13 heart disease characteristics.

### 4.2. Data Preparation

This includes cleaning, transformation, and normalizing the data before its usage in the model-building process. The body mass index features (BMI) were established using known characteristics such as height and weight.

### 4.3. Data Preprocessing

The datasets are pre-processed subsequently data cleaning, and the given dataset is segregated into training and test sets. Any machine learning method requires data pre-processing to produce relevant results. Most researchers utilize 70 % for training and 30% for testing, as stated by [K. Korjus, M. N. Hebart, and R. Vicente (2016)] because the system generates the more data dedicated to training, the more ideal and accurate outcomes. As a result, the 70:30 partitioning ratio is used.

### 4.4. Feature Selection

In a Machine Learning system feature selection is a technique that reduces the size of data by taking away the irrelevant and redundant features. The advantages of feature selection methods include model interpretability, quick training computation time, and reduced overfitting in classification models due to increased generalization. Before classification, the filter and wrapper method is used to choose a subset of selective characteristics by eliminating attributes that have little or no impact.

#### 4.4.1 Chi-square test

To evaluate features the filter evaluation criteria directly employ the statistical properties of the novel dataset. The evaluation process brings out improved computing efficiency [Talavera L.(2005)]. A Chi-Square method is a standard suggested approach for categorical variables. . Chi-square test is used to calculate that whether each feature value in the dataset is expressively different from the target value. The number of features is selected based on scores.

#### 4.4.2 Recursive Feature Elimination

It's a wrapper technique for selecting features. It removes redundant and weak features that have little impact on the pieces of training error while preserving the liberated and robust features to increase the model's simplification performance. It employs a method known as iterative feature _ranking. Feature ranking has been performed in the form of the backward feature elimination method. The procedure began with the development of a model was based on the whole set of features, followed by a rating of the relevance of each feature. After

removing the least significant feature, the models are rebuilt, and the importance of the features may be reviewed. [Mathew, T. E. (2019)].

### 4.5. Performance Evaluation Metrics

The major objective of this study is to compare machine learning models for cardiovascular disease prediction. Classification accuracy, f1_score, sensitivity, specificity and precision were used to assess the performance of model and it also chose the best model. With the help of the confusion matrix, these metrics are calculated. The performance evaluation metrics were calculated using Table 1.

| | Predicted 0 | Predicted 1 |
|---|---|---|
| 0 | TP | FP |
| 1 | FN | TN |

**Table 1 – Confusion Matrix**

Accuracy
The accuracy calculation formula was devised by
Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$ -- (1)

Precision
Precision refers to the ability to be precise and correct. Precision conveys the impression of accurately foreseen events. It calculates the percentage of positives that are genuine.

Precision = $\frac{TP}{TP+FP}$ -- (2)

Sensitivity:
It's the proportion of newly diagnosed heart patients to all heart patients. The "true positive rate" pertains to the sensitivity of the classifier in identifying positive cases. In other words, if a diagnostic test indicates that a person has a condition, the sensitivity (true positive fraction) confirms it. It is also called as recall

Sensitivity= $\frac{TP}{TP+FN}$ -- (3)
Specificity:
The proportion of those who had a negative result on this test out of those who do not have the disease is referred to as specificity (True Negative Rate). The total specificity is the percentage of correctly detected projected negative situations and it can be expressed mathematically as:

Specificity = $\frac{TN}{TN+FP}$ -- (4)

F1_score:

The F1_score, is a measurement of a model's accuracy on a particular dataset. It's used to evaluate binary classification algorithms that categorize examples into 'positive' and 'negative' groups. The F-score is a means of combining the precision and recall of a model.

F1_score= $2 \times \frac{Precision \times Recall}{Precision + Recall}$ -- (5)

## 5. Experiment Results and Analysis

The study's main goal is to evaluate the performance of four machine learning algorithms: Decision Tree, Logistic Regression, Random Forest, and Ada Boost. The data was processed in this study by calculating BMI based on heights and weights. The power of filter and wrapper feature selection methods was used in the classification model. The data is disunited into two sets: training and testing, with a ratio 70:30. In the first phase, chi-squared, filter methods are employed in this work. Since the chi-square test assumes a frequency distribution and does not accept negative values, non-negative data was removed from the dataset. The missing value will be overcome by the KNN imputer method. In the second phase, Recursive Feature Elimination was used. Feature selection methods discussed above evaluate the importance of the features by evaluating the importance of the filter method with reference to the class label for each feature in the dataset. The weights of features that met the precise condition in respect to the weights of input features were chosen for the datasets. The most popular metrics were used to measure the efficacy of machine learning algorithms.

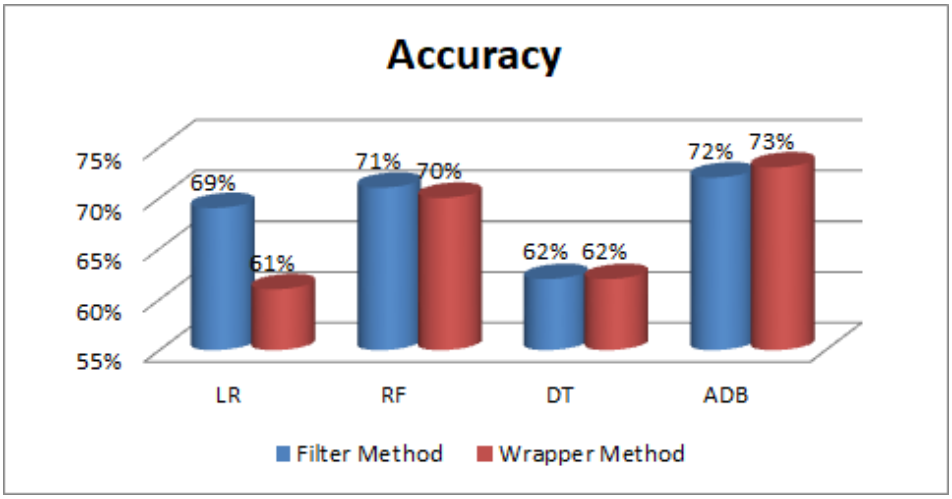


**Figure 2 Comparison of Accuracy**

Figure 2 compares the performance accuracy of the various strategies. When the two FS approaches are compared, the Ada Boost Classifier delivers the best results. LR produces lowest accuracy when compared with other model in Filter method and DT LR produces lowest accuracy in wrapper method. When compared to other algorithms, LR has the lowest accuracy in the Filter method, and DT has the lowest accuracy in the Wrapper method.

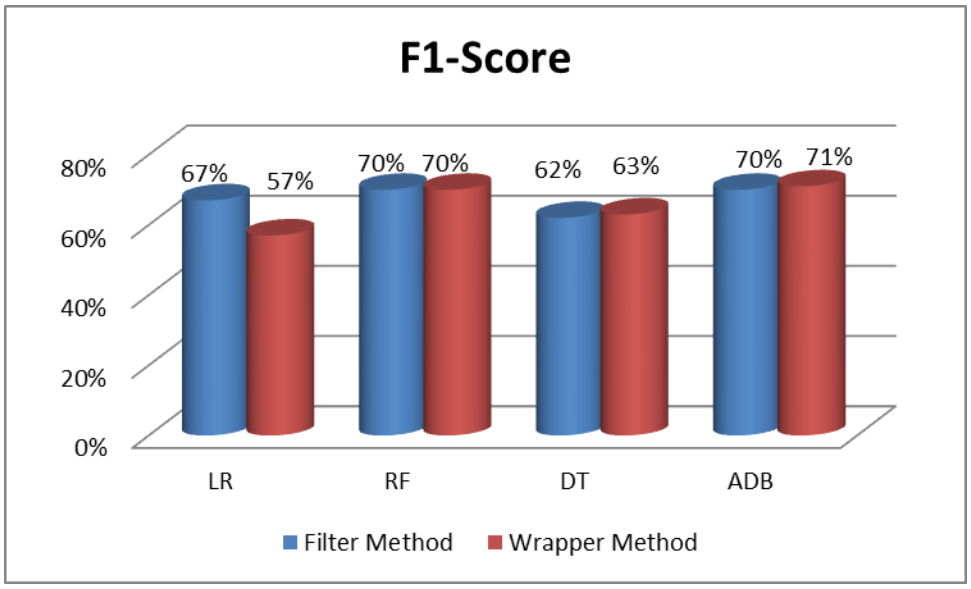


**Figure 3 Comparison of F1_Score**

Figure 3 demonstrates that the Ada boost classifier scored 70 % F1_score in the filter method and 71 % F1_score in the wrapper technique, the highest among the models.

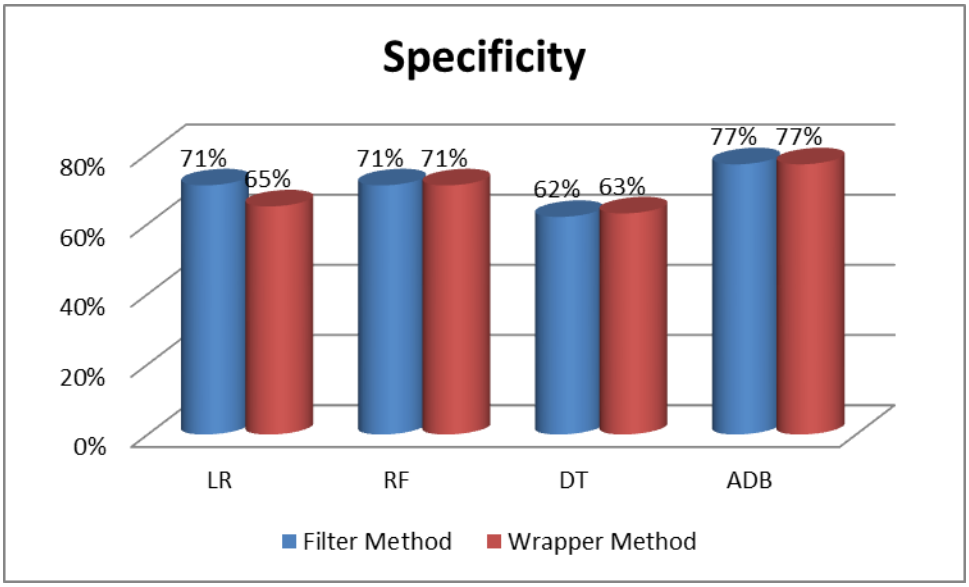


**Figure 4 Comparison Specificity**

The percentage of people who are correctly excluded by a test despite not having the condition is shown in Fig. 4. In both feature selection strategies, the classifier Ada boost produced the highest result of 77 %.

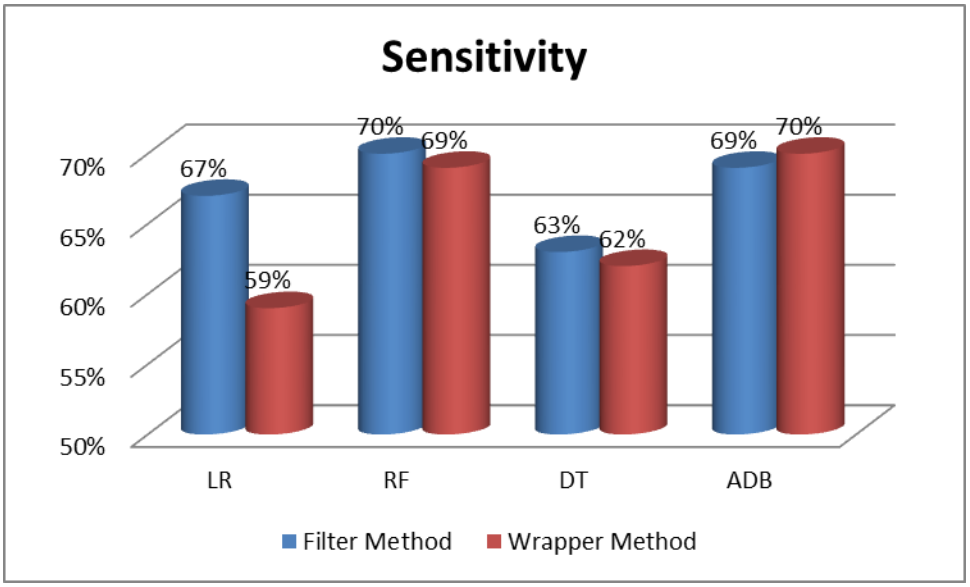


**Figure 5 Comparison of Sensitivity**

The proportion of predicted positive classes produced from whole positive examples from the dataset is given in Fig. 4, along with a comparison of the algorithms' recall measures or sensitivity.

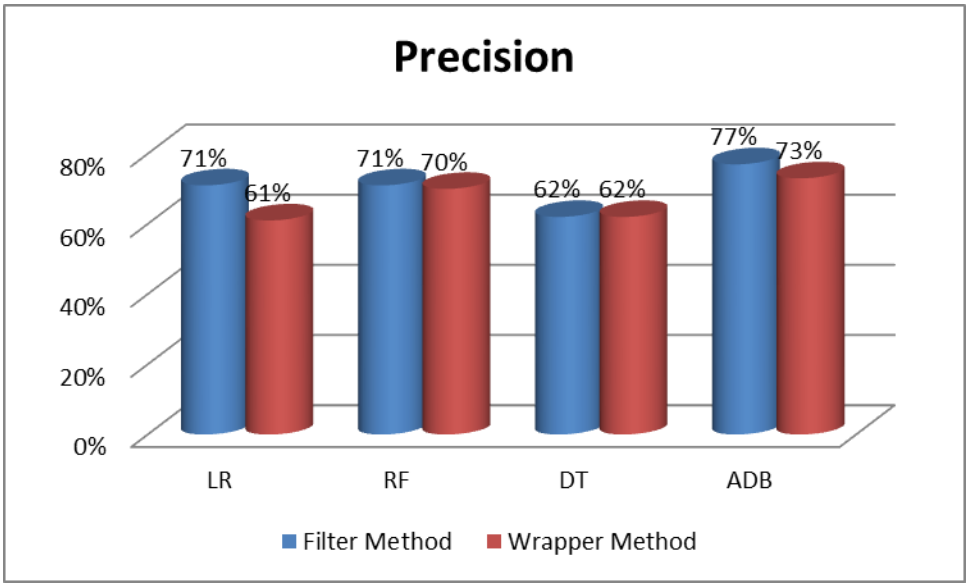


**Figure 6 Comparison of Precision**

The precision comparison is depicted in Figure 6, which determines the number of positive class predictions that substantially fit the positive class. In both Feature methods, the Ada Boost outperforms other algorithms by 77%. The methodology proposed compares the best

CHI and RFE outcomes. The performance measures used to evaluate the models are presented in tables below.

| Metrics | LR | RF | DT | ADB |
|---|---|---|---|---|
| **Accuracy** | 69% | 71% | 62% | 72% |
| | 61% | 70% | 62% | 73% |
| **F1_Score** | 67% | 70% | 62% | 70% |
| | 57% | 70% | 63% | 71% |
| **Precision** | 71% | 71% | 62% | 77% |
| | 61% | 70% | 62% | 73% |
| **Sensitivity** | 67% | 70% | 63% | 69% |
| | 59% | 69% | 62% | 70% |
| **Specificity** | 71% | 71% | 62% | 77% |
| | 65% | 71% | 63% | 77% |

**Table 2 Classifier Results.**

From the Table 2 LR is logistic regression, RF is random forest, DT is decision tree and ADB is Ada boost. Metrics represent the evaluation of the model. The classification of Heart disease data was one of the study's key approaches. This work contributes to identifying Ada Boost (ADB) as the best technique for predicting the occurrence of heart disease with improved accuracy for early diagnosis and also results in decreased death rates. As a result, the method benefits not only doctors but also patients by lowering laboratory examination costs and saving time.

## 6. Conclusion

An effective machine learning-based approach for analysis of heart disease was established in this work. The work was designed with the help of machine learning classifiers such as Decision Tree, Logistic Regression, Random Forest, and Ada Boost. The dataset used in the study comprises a number of patients affected by heart disease and also includes related features to perform the prediction. The prevalence of features in this dataset was determined by filter and wrapper feature selection procedures. These are the methods that were used to resolve the issue. Relevant features are used in the classifier model to perform evaluation metrics. Accuracy, f1 score, precision, sensitivity, and specificity performance evaluation measures were also used to evaluate the identification system's performance. According to Table 2, the Ada Boost classifier produces the best results in both feature selection algorithms when compared with other classifiers. When compared to other ways of filter feature selection, wrappers produce better results. Furthermore, irrelevant features harm the diagnostic system's performance and lengthen computation time. As a result, another ground-breaking part of this research was the use of feature selection algorithms to determine the

appropriate features, which enhanced classification accuracy while simultaneously reducing the computation of the diagnosis process. Other feature selection techniques and optimization approaches will be exploited to improve the performance of a prediction method for analyzing HD in the future.

## Reference

[1] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[2] Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA (2019) An automated diagnostic system for heart disease prediction based on Chi square statistical model and optimally configured deep neural network. IEEE Access 7:34938–34945.https ://doi.org/10.1109/ ACCES S.2019.2904800

[3] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146

[4]Breiman L., 2001. Random forests”, Mach Learn, 45(1):5–32.

[5] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, 100330.

[6] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013

[7]Jabbar MA, Deekshatulu BL, Chandra P (2016) Prediction of heart disease using random forest and feature subset selection. In: Innovations in bio-inspired computing and applications. Springer, Cham, pp 187–196. https ://doi.org/10.1007/978-3-319-28031 -8_16

[8] K. Korjus, M. N. Hebart, and R. Vicente, “An efficient data partitioning to improve classification performance while keeping parameters interpretable,” PloS one, vol. 11, no. 8, p. e0161788, 2016.

[9] Khourdifi Y, Bahaj M. Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. Int J Intell Eng Syst. 2019;12(1):242–52. https://doi.org/10.22266/ijies2019.0228.24

[10] Kim JK, Kang S (2017) Neural network-based coronary heart dis- ease risk prediction using feature correlation analysis. J Healthc Eng 2017:1–13. https ://doi.org/10.1155/2017/27805 01

[11] M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propa- gation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[12] Mathew, T. E. (2019). A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. 10(3), 55–63.

[13] Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.

[14] Rath, M.; Pattanayak, B.(2019). Technological improvement in modern health care applications using Internet of Things (IoT) and proposal of novel health care approach. International Journal of Human Rights in Healthcare, 12: 148–162. doi: 10.1108/IJHRH-01-2018-0007.

[15] Ruan, Y., Guo, Y., Zheng, Y. et al. Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1. BMC Public Health 18, 778 (2018). https://doi.org/10.1186/s12889-018-5653-9

[16] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, ―Improving Heart Disease Prediction Using Feature Selection Approaches,‖ in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST),

[17] S.Ulianova, "Cardiovascular Disease dataset," Kaggle.com, 2019. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset (accessed Jan. 02, 2021).

[18] Singh, Y., Sinha, N., Sanjay, S.: Heart disease prediction system using random forest, pp. 613– 623 (2017). https://doi.org/10.1007/978-981-10-5427-3_63

[19] Talavera L, An evaluation of ¯lter and wrapper methods for feature selection in categorical clustering, Lect Notes Computer Sci 3646:440, 2005.

[20] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical informatics and Decision Making, 19(1). https://doi.org/10.1186/s12911-019-1004-8

[21] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ‘‘A hybrid classification system for heart disease diagnosis based on the RFRS method,’’ Comput. Math. Methods Med., vol. 2017, pp. 1–11, Jan. 2017.