PREDICTION OF CARDIOVASCULAR DISEASES USING MACHINE LEARNING ALGORITHMS

1. Abstract

The heart is the most vital organ of any living being. It gives life and ensures blood reaches across the body. And yet heart diseases are the leading cause of death these days. Heart Diseases not only involve strokes and heart attacks but also lesser known yet deadly diseases like (i) Aortic stenosis - a condition in which the aortic valve, which regulates blood flow out of the heart, becomes narrowed, making it difficult for blood to flow through; (ii) Pulmonary hypertension - high blood pressure in the pulmonary artery, which can cause the right side of the heart to become enlarged, leading to heart failure; and Atrial fibrillation, Mitral valve prolapse, Hypertrophic cardiomyopathy, etc. What makes these diseases more deadly is the fact that these are asymptomatic diseases, i.e, they may not show the symptoms until it is too late. Hence requiring regular check-ups from doctors. But even still there is a very high chance of error in prediction the disease accurately due to factors such as (i) expensive tests that most people can't afford, (ii) Misinterpretation of test results, (iii) Failure to consider the patient's previous symptoms and history, which could be due to patients not disclosing shameful history with doctors. Many of the errors can be narrowed down to the possibility of human error, which is what we are trying to eliminate through this paper. Machine Learning algorithms have a significant role to play in the future in the prediction of heart disease as various models can be trained to predict heart diseases with very high accuracy. In this paper, we are going to see how we can improve the efficiency of predicting heart diseases. We are going to see what effect using cross ross validation and feature selection methods are going to have on our accuracy.

Keywords: Machine Learning Algorithms, Heart Diseases, SVM, Decision Tree, Random Forest, Neural Networks, Fischer Score, Recursive Feature Elimination, Cross Validation

2. Introduction

Heart Diseases can easily be identified and predicted using machine-learning Algorithms. They provide many advantages such as (i) High accuracy; (ii) Handling large and complex data – which might include rare and specialized conditions for individual patients; (iii) Automation of predictions- thereby reducing human error and being able to analyze data for multiple patients simultaneously, etc. [21]. In this paper, we are going to analyze the efficiency of machine learning algorithms like SVM, Decision Tree, Random Forests, and Neural Networks in predicting heart diseases using the confusion matrix, while also analyzing what impact cross validation and feature selection have on them [22]. Further, we shall analyze what type of feature selection is the most ideal for this task – wrapper or filter [23]. In total we will be analyzing 24 combinations and by the end of this paper we would be able to See the best combination for predicting heart diseases.[24]

3. Literature Survey

The paper describes a machine learning model that utilizes the Navis Bayesian algorithm. The model was trained and tested on the UCI database, using an 80-20 split for train-test. The model achieved an accuracy of 89.77% and incorporated other techniques such as Sequential Minimal Optimization (SMO) which had an accuracy of 84.07%, Bayes Nest with an accuracy of 81.11%, and Multi-layer Perception with an accuracy of 77.4%. Additionally, the model was implemented using WEKA for data mining and AES for security. [1]

The authors used a combination of five algorithms, Support Vector Machine (SVM), Decision Tree, Naive Bayes, K-Nearest Neighbour, and Random Forest to develop their model. They used the UCI database to train and test the model. The model used 14 attributes and the highest accuracy was achieved by the Random

Forest algorithm with an accuracy of 0.890110. The model also indicated the disease presence (1,2,3,4) and absence (0) numerically.[2]

A combination of three algorithms, Decision Tree, Naive Bayes, and Neural Network is used to develop the model. They used the UCI database with 14 attributes and 668 records to train and test the model. The highest accuracy was achieved by Decision Tree algorithm with an accuracy of 98.54%. The author noticed that sometimes removing features that are independent can result in higher accuracy.[3]

The Multilayer Perceptron Neural Network (MLPNN) with Back Propagation (BP) algorithm is used to train the model. The train-test split used was 40-60. The model achieved 100% accuracy in the training dataset. The authors also implemented the model using WEKA for data mining and developed an effective Heart Disease Prediction System (EHDPS). [4]

The authors of the paper proposed a Hybrid Algorithm which incorporates K-mean clustering algorithm and apriori algorithm for disease prediction using ECG signals. The authors used a centrality and frequency measure to feed each attribute individually into the K-means clustering algorithm which then clusters the information into similar groups. This information is then used to categorize the attributes/factors used in the disease prediction. [5]

Proposed model reportedly achieved an accuracy rate of 96.7% on the test set, which was trained using seven different characteristics. To construct the decision tree component of the model, the authors employed the Classification and Regression Tree (CART) algorithm. [6]

Used a combination of five machine learning algorithms, Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest Classifier, and Support Vector Machine to develop the model. It used a 90-10 train-test split, the dataset consists of 14 attributes and 303 instances. The highest accuracy was achieved by Random Forest Classifier and Support Vector Machine algorithm with an accuracy of 90.32%.[7]

The model was created using the WEKA data mining tool and the UCI dataset. The performance metric, such as accuracy and train-test split, as well as the technique utilised, were not disclosed by the authors. They mentioned improving algorithm accuracy by using a hybrid technique.[8]

A combination of the Naive Bayes Algorithm, Logistic Regression, and Random Forest Algorithm was used to develop the model. It used the Kaggle heart disease prediction dataset and a 70-30 train-test split. The highest accuracy was achieved by the Random Forest algorithm with an accuracy of 99%. Logistic Regression and the Naive Bayes algorithm achieved an accuracy of 75% and 87% respectively.[9]

A combination of Neural Networks, Fuzzy Logic and Decision Tree to develop the model. The model used 7 attributes Family History of Heart Disease, Smoking, Cholesterol, High Blood pressure, Obesity, and Lack of physical Exercise. The highest accuracy was achieved by the Neural Networks algorithm with an accuracy of 100% and the Decision Tree algorithm with an accuracy of 99.62%. The authors used WEKA as the data mining tool and proposed developing an Intelligent Heart Disease Prediction System (IHDPS).[10]

The authors used a combination of Neural Networks, SVM, and KNN algorithms to predict heart disease using a dataset of Algerian patients collected from the Mohand Amokrane EHS Hospital. The model used 20 attributes and the highest accuracy was achieved by the Neural Networks algorithm with an accuracy of 93%. The authors also used a confusion matrix to test the algorithms. The goal of the study was to create a reliable and accurate system for identifying heart disease patients and providing early treatment.[11]

The authors of the paper used decision tree algorithms, J48 and Random Forest, to predict heart disease using UCI database. The Random Forest algorithm was found to be more accurate than J48. The data mining tool WEKA was used in the study.[12]

The authors of the paper compares the accuracy of several machine learning algorithms, including Navis Bayes, Support Vector Machine, K-NN, Decision Tree, and Random Forest, for predicting heart disease using the People's Hospital Database. The Random Forest algorithm was found to have the highest accuracy at 91.6%, significantly higher than the other methods. The key features of the dataset were selected by using the CFS(Correlation-based Feature Selection) Subset Evaluation method in combination with Best First Search to reduce dimensionality.[13]

The authors of the paper presents an enhanced algorithm called ENDDP which is used to predict heart disease using the UCI database. The study compares the accuracy of the Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF) classification models. 14 out of 76 attributes were selected for the analysis and the ENDDP algorithm was used to obtain better results. The presence or absence of the disease is indicated numerically (1,2,3,4 for presence and 0 for absence) in the dataset.[14]

The authors of the paper uses trained recurrent fuzzy neural networks (RFNN) to diagnose heart disease using the Cleveland dataset. The study uses 14 out of 76 attributes selected by Correlation feature selection method. The study found that Correlation feature selection method in RFNN has the best operation in feature selection with an accuracy of 98.4%. Additionally, the study compares five kinds of feature selection and extraction models such as data envelopment analysis (DEA), Linear Discriminative Analysis (LDA), Principle Component Analysis (PCA), Correlation Feature Selection (CFS), and Relief. No information is provided on the train-test split, or any additional points.[15]

The study used a combination of the Naive Bayes algorithm and a genetic algorithm to predict heart disease. The data used came from the UCI database and included 14 attributes such as age, gender, resting blood pressure, cholesterol, fasting blood sugar, and old peak. MySQL database was used.[16]

The study used 3 Bayes classifiers algorithms (Naive Bayes, Bayes Net, and Naive Bayes Multinomial) to predict heart disease. The authors used 14 attributes and 271 instances to make predictions. The accuracy was measured using MAE, RMSE, RAE, and RRSE, with Naive Bayes Multinomial having the highest accuracy. The study used the cross-validation method and Data mining tool Weka, in the test option there are four kinds of parameters like supplied test set, training set, percentage split, and cross-validation.[17]

The study used Naive Bayesian and Support Vector Machine (SVM) algorithms to predict heart disease. The data used in the study came from the UCI database, with a 75-25 train-test split and 15 attributes, and 303 records. The study found that SVM had higher accuracy (55%) compared to Naive Bayes (40%).[18]

In this study, a combination of the Naive Bayes Classifier and K-Nearest Neighbor (cNK) algorithm was used to predict heart disease. The data used in the study came from the UCI database, with a 90-10 train-test split. The study aimed to combine the strengths of both algorithms to improve the overall accuracy of the model. The Naive Bayes classifier is simple, easy to implement, and requires fewer data and computational power, while the KNN algorithm is instance-based and stores all the available cases. The study found that cNK algorithm had the highest accuracy of 85.92%.[19]

In this study, a combination of Neural Network and Genetic Algorithm was used to predict heart disease. The data used in the study came from the Z-Alizadeh Sani dataset, with a 90-10 train-test split and 54 attributes. The study aimed to utilize the strengths of both algorithms to improve the overall accuracy of the model by using a neural network for pattern recognition and a genetic algorithm for optimization. The study used RMSE to evaluate the model and used feature selection algorithms such as the Gini index, weight by SVM, information gain, and principal component analysis (PCA) to pre-process the data before feeding it to the model.[20]

4. Data Set Used

For this research, we use the UCI (UC Irvine) Cleaveland Dataset. It contains attributes and details of 297 patients, some of which had and others who didn't have heart disease. A total of 14 attributes were used for research. We are going to implement a 80:20 train test split for this research.

14 attributes used:
Age; Sex; Chest Pain; Resting blood pressure; Cholesterol measurement; Fasting blood sugar; Resting electrocardiographic results; Maximum heart rate achieved; Exercise induced angina; ST depression induced; Slope of the peak exercise; Number of major vessels; Thalassemia value; heart disease(target).

5. Algorithms Used

1. **Prediction Algorithms:**
   - SVM (Support Machine Vector)
   - Decision Tree
   - Random Forest
   - Neural Network
2. **Feature Selection:**
   - Wrapper method - Recursive Feature Elimination
   - Filter Method - Fischer Score
3. **Cross Validation**
4. **Accuracy – Confusion Matrix**

A. Prediction Algorithms

I. **SVM (Support Machine Vector):**

Support Vector Machines (SVM) are a type of supervised learning algorithm that can be used for classification and regression tasks. They work by finding the optimal boundary, called a hyperplane, that separates the data into different classes. The points closest to the hyperplane, called support vectors, have the greatest impact on the position of the hyperplane. SVMs can also handle non-linearly separable data by using kernels [26], which map the data into a higher dimensional space where a linear boundary can be found. The goal of an SVM is to find the hyperplane that maximizes the margin, or the distance between the hyperplane and the closest data points from each class. [34]

Using Kernels in SVM can help improve the efficiency of the model in predicting correctly by being able to evaluate multiple dimensions of data.

II. **Decision Tree**

The Decision Tree algorithm is a supervised learning method commonly used in classification problems. It is adept at handling both continuous and categorical attributes. The algorithm partitions the data into subsets by selecting the most significant predictors. The process begins by calculating the entropy of each attribute, then using this information to determine the best split point using the variable or predictor with the highest information gain or lowest entropy. These steps are repeated recursively with the remaining attributes. The result is a tree-like structure that represents the decisions and predictions made based on the input data. Formulas such as entropy and information gain are used to determine the best split point at each step in the decision tree algorithm.[35]

III. **Random Forest:**

Random Forest is a very powerful supervised machine learning algorithm that can be used for both classification and regression problems. It is an ensemble method that creates multiple decision trees and combines the predictions of all the trees to make a final prediction.[29]

IV. **Neural Network**

Neural network imitates the structure and operation of the human brain. They are extensively utilised in many different applications, including as prediction, natural language processing, and image recognition. They are made up of interconnecting layers of artificial neurons that have been trained to spot patterns in data. In order to reduce the error between the expected output and the actual output, the neurons' weights and biases are changed throughout training. The activation function, which controls a neuron's output, and the backpropagation algorithm, which updates the weights and biases during training, are two essential formulae used in neural networks.[36]

B. Feature Selection

I. Wrapper Method

In the wrapper technique, features are chosen by treating them as a search problem, where many combinations are created, assessed, and contrasted with other combinations. Iteratively employing the subset of characteristics trains the algorithm.

For this research we are going to use Recursive feature elimination and Backward feature elimination for implementing the wrapper method of feature selection.

Recursive feature elimination (RFE) is a method used to select features for a given external estimator. In this method, though we start with all the attributes, the least significant attributes are removed after each iteration. The importance of the attribute is typically measured using performance metrics such as F1-Score.[27]

Whereas Backward Feature Elimination (BFE) is a feature selection method that starts with all the attributes and eliminates the most important attributes one by one. [25]

II. Filter Method

In the filter technique of feature selection, relevant features are identified and chosen as a pre-processing step independent of the learning algorithm using statistical measurements. Through the ranking of features according to several criteria, this strategy eliminates duplicate and superfluous columns from the model. Filter techniques provide the advantages of quick calculation and less chance of overfitting the data.

For this research we are going to use Fisher Score for implementing the wrapper method of feature selection.[31]

The Fisher score method provides a score to each feature depending on the discriminating power of the feature. The user may choose the top-scoring features to include in their model by sorting the features according to the algorithm, which returns them in descending order of their Fisher scores. The Fisher score has the advantage of being independent of the learning algorithm being applied, which makes it a flexible feature selection strategy. Fisher score also does not overfit the data and is computationally efficient.[32]

C. Cross Validation

During Cross Validation, the data is divided into numerous subsets, or "folds," and the algorithm is trained on each subset before being evaluated on the entire data. As a result, the algorithm's performance may be evaluated more thoroughly since it is tested on a wide range of data samples. Cross-validation can increase the effectiveness of machine learning algorithms by lowering the risk of overfitting and giving a more precise assessment of the algorithm's performance on unknown data by averaging the results from numerous iterations of this procedure. Additionally, the algorithm's parameters may be adjusted via cross-validation to enhance performance.[33]

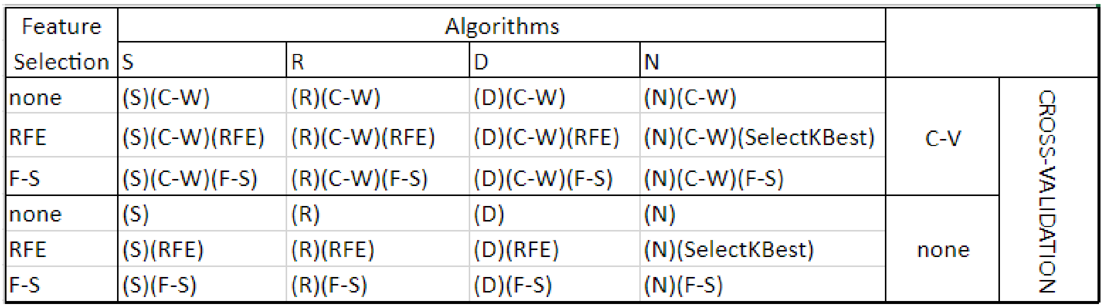
During this research, we are going to use a 1/3 split of dataset for cross validation.

D. Confusion Matrix

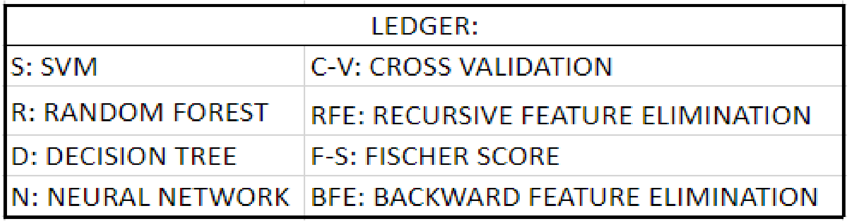
Confusion Matrix is a method to evaluate the efficiency of a program. Original and Predicted values are passed onto the algorithm and it returns the efficiency in a tabular format with attributes such as True Positive, True Negative, False Negative and False positive. These attributes are then used to calculate performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide a more comprehensive understanding of the model's performance than simply looking at the classification accuracy alone. [30]

6. Working

In order to do a comprehensive analysis of all possible variations we are going to create 24 separate variations, which are as shown:



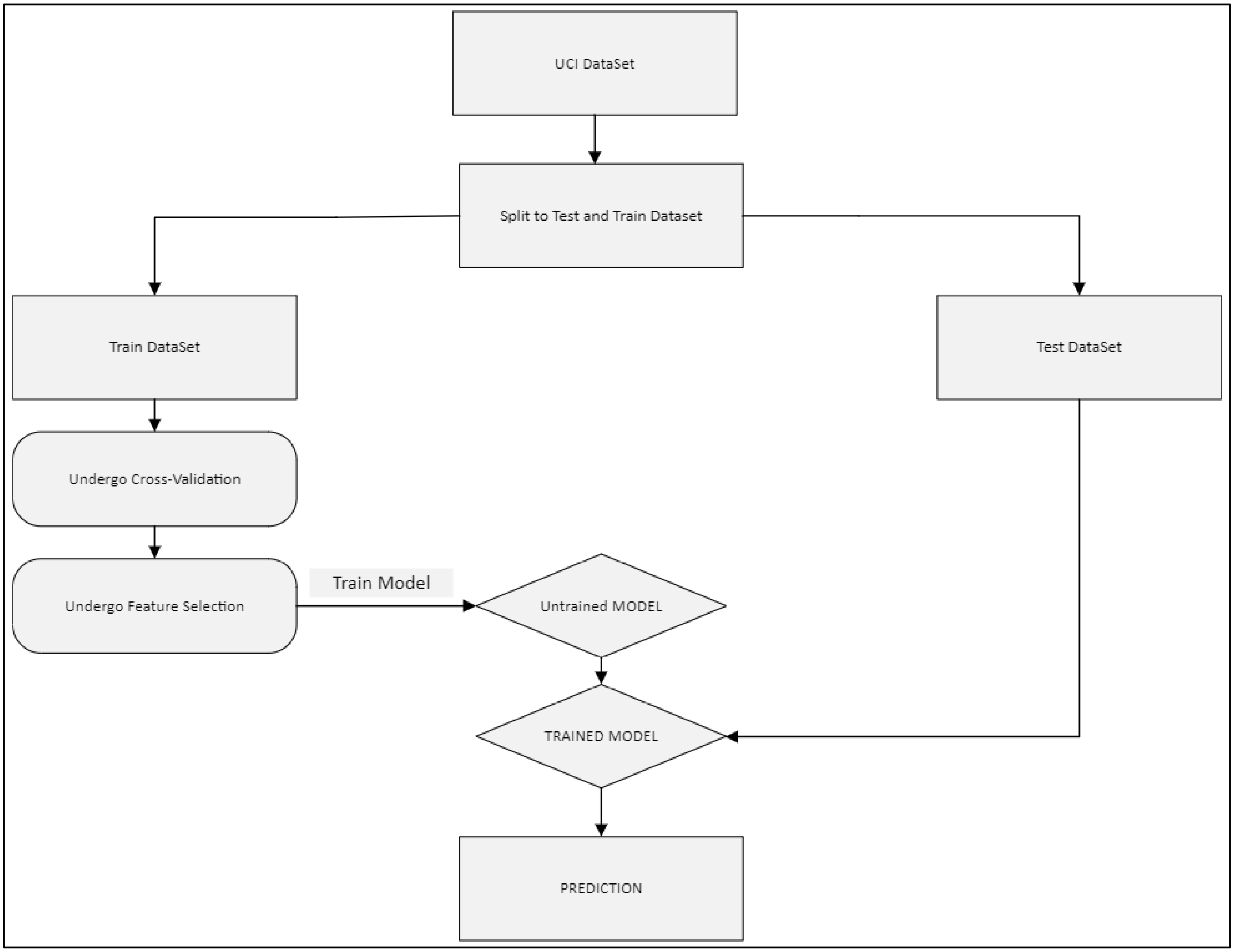
Feature Selection	Algorithms					
	S	R	D	N		
none	(S)(C-W)	(R)(C-W)	(D)(C-W)	(N)(C-W)	C-V	CROSS-VALIDATION
RFE	(S)(C-W)(RFE)	(R)(C-W)(RFE)	(D)(C-W)(RFE)	(N)(C-W)(SelectKBest)		
F-S	(S)(C-W)(F-S)	(R)(C-W)(F-S)	(D)(C-W)(F-S)	(N)(C-W)(F-S)		
none	(S)	(R)	(D)	(N)	none	
RFE	(S)(RFE)	(R)(RFE)	(D)(RFE)	(N)(SelectKBest)		
F-S	(S)(F-S)	(R)(F-S)	(D)(F-S)	(N)(F-S)		



LEDGER:	
S: SVM	C-V: CROSS VALIDATION
R: RANDOM FOREST	RFE: RECURSIVE FEATURE ELIMINATION
D: DECISION TREE	F-S: FISCHER SCORE
N: NEURAL NETWORK	BFE: BACKWARD FEATURE ELIMINATION

Once the analysis of all 24 combinations is done via the confusion matrix. They are compared with each other for determining various conclusions:

1. When classifier algorithms are used without any additional algorithms, which one provides the best accuracy?
2. Which mode of feature selection provides the highest accuracy on average?
3. Which model is better suited to deal with other attributes than accuracy in the confusion matrix?
4. What effect is cross validation having on the accuracy of the models?
5. When comparing all 24 variations which model provides the best accuracy?



7. Result

During the execution of experiments, we came across a few observations:

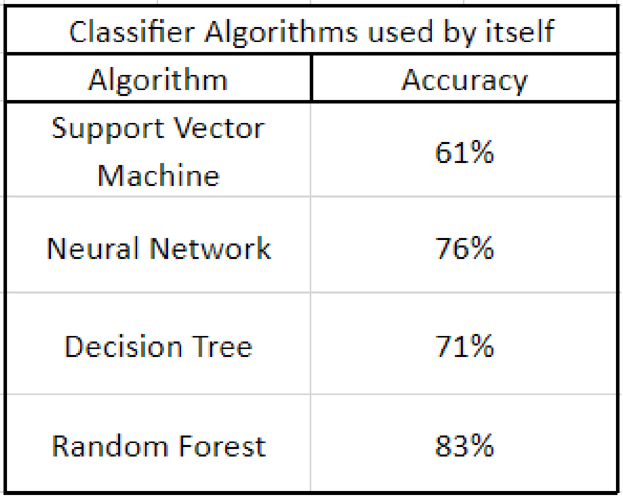
1. Recursive Feature Elimination could not be used for SVM and Neural Network.

This is because recursive feature elimination works by recursively removing the weakest attributes until a desired number of attributes is reached. It is mostly applicable for linear models such as linear regression and linearSVM as it's based on the idea that features with the lowest coefficients in a linear model are less important and are hence removed. Hence, they are not suitable for nonlinear SVM models and neural networks. [28]
Therefore, we used Backward Feature Elimination and SelectKBest instead.

2. Accuracy of Random Forest goes down when Feature Selection was also implemented.
Using feature selection on a complex model like Random Forest can result in overfitting. This would negatively impact its performance with unseen data. The loss of data due to feature removal may cause underfitting, thereby affecting the accuracy of prediction. [29]

Evaluation of Results for Intended Questions:

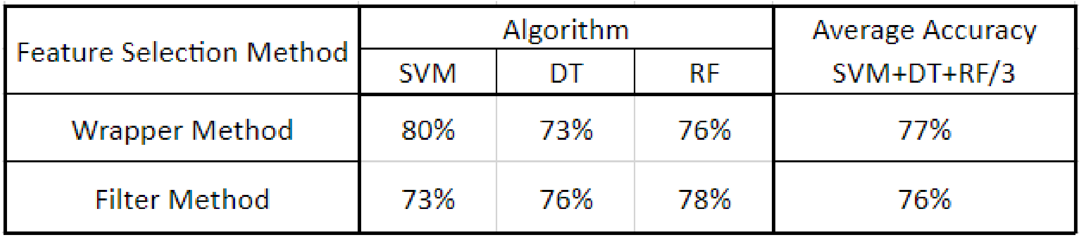
1. When classifier algorithms are used without any additional algorithms, which one provides the best accuracy?



Classifier Algorithms used by itself	
Algorithm	Accuracy
Support Vector Machine	61%
Neural Network	76%
Decision Tree	71%
Random Forest	83%

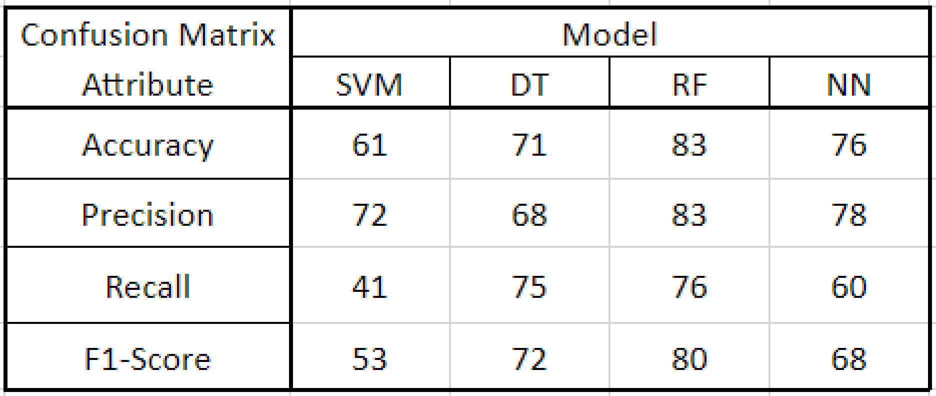
2. Which mode of feature selection provides the highest accuracy on average?

For this calculation, we did not use Neural Network because the wrapper method of feature selection could not be applied to neural network models.



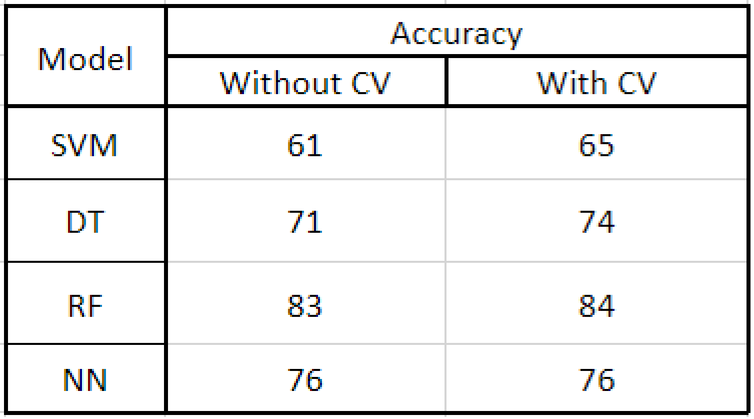
Feature Selection Method	Algorithm			Average Accuracy
	SVM	DT	RF	SVM+DT+RF/3
Wrapper Method	80%	73%	76%	77%
Filter Method	73%	76%	78%	76%

3. Which model is better suited to deal with other attributes than accuracy in the confusion matrix?



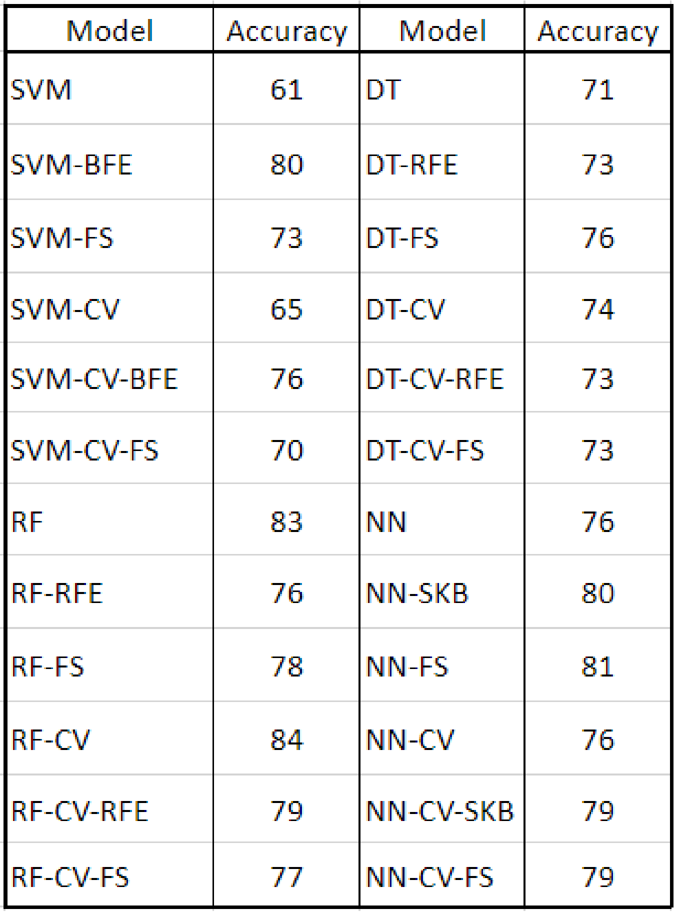
Confusion Matrix	Model			
Attribute	SVM	DT	RF	NN
Accuracy	61	71	83	76
Precision	72	68	83	78
Recall	41	75	76	60
F1-Score	53	72	80	68

4. What effect is cross-validation having on the accuracy of the models?



Model	Accuracy	
	Without CV	With CV
SVM	61	65
DT	71	74
RF	83	84
NN	76	76

5. When comparing all 24 variations which model provides the best accuracy?



Model	Accuracy	Model	Accuracy
SVM	61	DT	71
SVM-BFE	80	DT-RFE	73
SVM-FS	73	DT-FS	76
SVM-CV	65	DT-CV	74
SVM-CV-BFE	76	DT-CV-RFE	73
SVM-CV-FS	70	DT-CV-FS	73
RF	83	NN	76
RF-RFE	76	NN-SKB	80
RF-FS	78	NN-FS	81
RF-CV	84	NN-CV	76
RF-CV-RFE	79	NN-CV-SKB	79
RF-CV-FS	77	NN-CV-FS	79

8. Conclusion

Conclusions Derived from the research:

1. Random Forest (83%) provides the most accuracy, compared to other models when feature selection or cross-validation is not applied. The rest of the models follow the order of Neural Networks (76%), Decision Tree (71%), and SVM (61%). Where support vector machine provides the least accuracy.

2. When wrapper and Filter Methods were implemented on SVM, Decision trees, and Random Forest. Wrapper methods (RFE, BFE) were found to have a slightly higher accuracy (77%), compared to Filter methods (Fischer Score) (76%)

3. To have higher accuracy, Random Forest is the most suitable model. Random Forest was also found to be the best model to have higher precision, Recall and F1- Score.

4. Cross-validation increases the accuracy of the models, as it prevents the overfitting of models and provides more precision on unseen data. Thereby, increasing the accuracy.

5. The best combination of algorithms for higher accuracy was found to be Random Forest with cross-validation (83%). The worst accuracy of all was that of Support Vector Machines (61%).

9. Future Work

The efficiency of the prediction algorithms can be further improved by training them with multiple datasets. Also, more critical attributes such as family history related to the disease, consumption of junk food, use of substances such as alcohol and tobacco, and exercise must be considered. Around the world, there is still a shortage of doctors and tests are time-consuming and expensive. Hence, developers need to work with healthcare professionals to develop and make this kind of disease prediction technology publicly available.

10. Acknowledgment

I would like to thank my professor, Mrs. Arti Pawar, who guided me throughout this paper from the very beginning till the end. She had initially encouraged me into doing this research and had helped me whenever I felt stuck or didn’t know what to do next. Her deep knowledge in the field encouraged me to write a very good and comprehensive paper.

11. References:

1. A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, “Design and Implementing Heart Disease Prediction Using Naives Bayesian,” in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2019, pp. 292–297. doi: 10.1109/ICOEI.2019.8862604.
2. A. Chandra Patel *et al.*, “Prediction of Heart Disease Using Machine Learning,” *IJSDR1904075 International Journal of Scientific Development and Research*, 2019, [Online]. Available: www.ijsdr.org
3. Pratiksha Shetgaonkar and Dr. Shailendra Aswale, “Heart Disease Prediction using Data Mining Techniques,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 2, pp. 281–286, Feb. 221AD, [Online]. Available: www.ijert.org
4. P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” *Int J Nanomedicine*, vol. 13, pp. 121–124, 2018, doi: 10.2147/IJN.S124998.

5. S. Sangwan and A. Tazeem Ahmad Khan, “REVIEW PAPER AUTOMATIC CONSOLE FOR DISEASE PREDICTION USING INTEGRATED MODULE OF A-PRIORI AND K-MEAN THROUGH ECG SIGNAL,” 2015. [Online]. Available: www.ijtre.com
6. S. Anusuya and M. S. Barath, “Early Heart Disease Prediction Using Machine Learning,” 2022. [Online]. Available: www.ijrpr.com
7. P. Kumar Bhunia, P. Mondal, K. Ganguly, A. Debnath, M. D. E, and P. Rakshit, “Heart Disease Prediction using Machine Learning,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 11, pp. 272–274, 2021, [Online]. Available: www.ijert.org
8. Syed Ahmed Yasin and D. Rao, “Analysis Of Single And Hybrid Data Mining Techniques For Prediction Of Heart Disease Using Real Time Dataset,” 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET
9. Dr. Loganathan R, Syed Farooq, Sayeeda Arshiya, Supreksha Karki, and Syed Sohail, “Machine Learning Based Heart Disease Prediction System,” *Int J Sci Res Sci Eng Technol*, pp. 202–206, Feb. 2022, doi: 10.32628/ijsrset218543.
10. B.Kaur and W. Singh, “Review on Heart Disease Prediction System using Data Mining Techniques, *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 10, pp. 3003–3008, Oct. 2014, [Online]. Available: http://www.ijritcc.org
11. D. Upadhyay, A. Singh Dhabariya, H. Bohra, and B. Shrimali, “PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS AND NEURAL NETWORK,” *Online International Journal of Technical Research & Science*, vol. 7, no. 8, pp. 2454–2024, Aug. 2022, doi: 10.30780/IJTRS.V07.I08.001.
12. S.Spino, Dr.M.Mohamed Sathik, and Dr.S.ShajunNisha, “Prediction of Heart Disease using Decision Tree Classification Algorithms,” *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE)*, no. Special Issue ICRTETM March 2019, pp. 18–22, 2019.
13. T. Katari and B. MadhavaRao, “Machine Learning Algorithms for Predicting Heart Disease,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 8, no. 1, pp. 301–306, Feb. 2022, doi: 10.32628/IJSRCSEIT.
14. J. N. Rao and R. S. Prasad, “An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm for Predicting Heart Disease in Machine Learning,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 94–104, Oct. 2021, doi: 10.32628/cseit206429.
15. S. Kordnoori, H. Mostafaei, M. Rostamy-Malkhalifeh, and M. Ostadrahimi, “Diagnosis of Heart Disease Using Feature Selection Methods Based On Recurrent Fuzzy Neural Networks,” *IPTEK The Journal for Technology and Science*, vol. 32, no. 2, p. 64, Sep. 2021, doi: 10.12962/j20882033.v32i2.7075.
16. A. Xavier, S. Chakalakal, and S. Sadat, “Heart Disease Prediction using Machine learning and Data Mining Technique,” *International Journal of Engineering Research & Technology*, vol. 9, no. 3, pp. 303–305, 2020, [Online]. Available: www.ijert.org
17. R.Nithya, D.Ramyachitra, and P.Manikandan, “An Efficient Bayes Classifiers Algorithm on 10-fold Cross Validation for Heart Disease Dataset,” *International Journal of Computational Intelligence and Informatics*, vol. 5, no. 3, pp. 229–235, Dec. 2015.
18. P. Kanikar and D. Rajeshkumar Shah, “Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification,” *Int J Comput Appl*, vol. 156, no. 2, pp. 975–8887, Dec. 2016.

19. E. Z. Ferdousy, Md. M. Islam, and M. A. Matin, "Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models," *Computer and Information Science*, vol. 6, no. 3, May 2013, doi: 10.5539/cis.v6n3p48.
20. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Comput Methods Programs Biomed*, vol. 141, pp. 19–26, Apr. 2017, doi: 10.1016/j.cmpb.2017.01.004.
21. R. Agarwal and V. S. Dixit, "Heart disease prediction using machine learning techniques," International Journal of Computer Applications, vol. 175, no. 8, pp. 1-6, 2017. DOI: 10.5120/ijca2017913294
22. J. Al-Jarrah and S. Al-Nimri, "Predicting Heart Diseases Using Machine Learning Algorithms," Journal of Medical Systems, vol. 41, no. 12, pp. 571, 2017. DOI: 10.1007/s10916-017-0705-z
23. X. Zhang, Y. Zhang, and J. Zhang, "Heart disease diagnosis with improved deep learning algorithms," Journal of Medical Systems, vol. 43, no. 2, pp. 98, 2019. DOI: 10.1007/s10916-018-1176-x
24. M. H. Aljohani, "Comparison of Machine Learning Algorithms for Predicting Heart Diseases," Journal of Healthcare Engineering, vol. 2018, pp. 1-14, 2018. DOI: 10.1155/2018/7190153
25. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344
26. Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*(3), 1171–1220. https://doi.org/10.1214/009053607000000677
27. Kohavi, R., & John, G. H. (n.d.). *Wrappers for feature subset selection*. http://robotics.stanford.edu/
28. Guyon, I., Weston, J., & Barnhill, S. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learining*, *46*, 389–422.
29. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688
30. Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.
31. Nirmala, S., & S.P, Dr. (2011). A study on Feature Selection Techniques in Bio-Informatics. *International Journal of Advanced Computer Science and Applications*, *2*(1). https://doi.org/10.14569/ijacsa.2011.020121
32. Wang, S., Li, D., Wei, Y., & Li, H. (2009). A feature selection method based on Fisher's discriminant ratio for text sentiment classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5854 LNCS*, 88–97. https://doi.org/10.1007/978-3-642-05250-7_10
33. Garbo, A., & German, B. J. (2019). Performance assessment of a cross-validation sampling strategy with active surrogate model selection. *Structural and Multidisciplinary Optimization*, *59*(6), 2257–2272. https://doi.org/10.1007/s00158-018-02190-7
34. Leopold, E. (2002). *Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?* (Vol. 46).
35. Paul E. (1989). *Incremental Induction of Decision Trees* (Vol. 4).
36. Sumijan, Windarto, A. P., Muhammad, A., & Budiharjo. (2016). Implementation of neural networks in predicting the understanding level of students subject. *International Journal of Software Engineering and Its Applications*, *10*(10), 189–204. https://doi.org/10.14257/ijseia.2016.10.10.18