HEART DISEASE PREDICTION USING CLASSIFICATION (NAIVE-BAYES)

Akansh Gupta[1], Lokesh Kumar[2]
Dr. Rachna Jain[3] , Ms. Preeti Nagrath[4]
Bharati Vidyapeeth's College of Engineering, Delhi[1,2,3,4]
[1] akansh.gupta1298@gmail.com , [2] krlokeshv99@gmail.com ,
[3] rachna.jain@bharatividyapeeth.edu, [4] preeti.nagrath@bharatividyapeeth.edu

Abstract
This paper aims towards a greater idea and utilization of machine learning in the medical sector. In this paper, comparative performances of six classification models are presented, when used on UCI's Cleveland Heart Disease Records to predict coronary artery disease(CAD). At first, all the 13 provided independent features were used to build the models. On comparing the accuracy of models, it was found that K-Nearest Neighbors(KNN), Support Vector Machine(SVM), Naive Bayes have expected and better performances. Thereafter, feature selection is applied to improve prediction accuracy. The Backward Elimination Method and Filter Method based on Pearson's Correlation Coefficient is used to choose major predicting features. The accuracy of models using all features and using features selected significantly enhanced the performance of Naive Bayes and Random Forest, while the other models didn't perform as expected. Naive Bayes produced an accuracy of 88.16% on the test set thereafter.

Keywords:
Naive-Bayes, Random Forest, Classification Model, Coronary Artery Disease, Cleveland Dataset

1. INTRODUCTION

Heart diseases are common but serious health issue in most of the countries nowadays. The improvement in technology is making us physically less active, and susceptible to diseases. They use their intellect to work on machines, build them, upgrade and maintain their functions.

It is sometimes the case that a disease has different forms requiring various treatments. Like, among patients with angiographic coronary artery disease (CAD), patients with triple-vessel or left main CAD need surgical coronary artery revascularization, on the other hand, patients with narrowing in single or double vessel may benefit from medical therapy [1]. Intuitively, a physician would first decide in case the patient has CAD, and, if so, then the patient needs surgery or not. Therefore, in the medical sector, machine learning algorithms can be used to recognize some severe diseases using the previous trends of symptoms.

Physical lethargy and malnutrition, overweight and corpulence, tobacco and substance abuse are amongst the dominant agents of cardiovascular disease as suggested by the Rochester's Medical centre. Alwan [2] conferred the research in World Health Organisation, where he elucidated the prevention of heart disease. The research showed that heart diseases are at the top among non-communicable disease which sums for ⅓ of mortality rate and 10 percent of the global disease trouble.
Also, it is not equitable, to correlate the accuracy of models and resolve the finest since the operation depends upon data [3]. Numerous studies relate data mining and statistical concepts to crack prediction queries. The comparative studies have mainly considered a particular data set or the same distribution of the dependent variable.

Anatomy of this paper is as follows: *Section 2* will apprise you about the backdrop of heart disease prediction systems. *Section 3* will expose you to the dataset used in this research, also data exploration and pre-processing techniques. *Section 4* is composed with classification algorithms of machine learning used: KNN, SVM, Logistic Regression, Decision Tree and Random Forest; Feature extraction techniques will also be analysed. *Section 5* will include the Results and Discourse of the research conducted. The paper is accomplished in *Section 6* by considering the future also.

The main encouragement of this research is to recommend a prediction system using the best classification model so it could help the medical experts to make successful decisions.
The highest mortality rate in numerous countries is due to heart disease. Any successful treatment is always attributed by an accurate and precise diagnosis. Therefore heart disease prediction systems based on machine learning algorithms assist in such cases to get the right results.

The evaluation of models was based on confusion matrix. Different Feature Extraction methods were used to select the most related attributes from given dataset to find an algorithm which performs best on the given dataset.

2. BACKGROUND

Many researches have been accompanied on the prediction of CAD. At every moment, technology is leaping over what it had attained earlier. Things which seemed impossible before are now easy with the advancement of technology. Prediction of the disease using machine learning has become a common practice.

Considerable studies have appeared that were concentrated on heart disease analysis. Distinct methods have been tried to solve the given problem and they acquired high classification accuracies. Here are some citations:

Detrano et al. [4] had outcomes which achieved acceptable classification efficiency of almost 77% on administering logistic regression algorithm along derived discriminant. Zheng Yao [5] enforced an advanced model, based on C4.5 upgraded the efficacy of attribute selection and partitioning criteria.. A study also revealed that the rules devised by the new model on C4.5 (R-C4.5) can give fair and appropriate explanations in healthcare fields. Resul Das [6] applies an approach that employs Statistical Analytics application called SA software for diagnosing heart disease (ANN-based method). Imran Kurt et al. compare effectiveness of logistic regression, CART, & ANN for predicting CAD [7].

Jabbar et al. [8] employed 'Principal Component Analysis' as a medium to trim the dimensionality of dataset and then used neural networks to produce better performance than traditional methods. John Gennari's [9] visionary CLASSIT system of clustering exhibited an accuracy of 78.9% on UCI's Heart Disease Data. Alfeo Sabay's [10] model based on neural network which worked upon *fabricated data* was capable to enhance the accuracy of CAD prediction to 96.7%. The 'Majority Voting Ensemble Method' used by Mechnovic yielded 87.37% for binary classification only outperformed by ANN [11].

3. PRELIMINARIES OF RESEARCH

3.1 HEART DISEASE DATASET

The following research has been performed using data from the Heart Disease Database available at the UC Irvine Repository [12]. This data has been available since 1988 and used by many researchers in heart disease prediction research because of its availability.
There are four available dataset, out of these four sources, only the Cleveland dataset has been used in machine learning experiments mostly, due to its completeness of observations. Well, it was repeatedly mentioned in previous studies but still exploring the dataset during the data preprocessing stage showed it has the only six observations of missing data attributed with '?'.

(Table 1.1: Data Type of attributes of Cleveland Dataset)

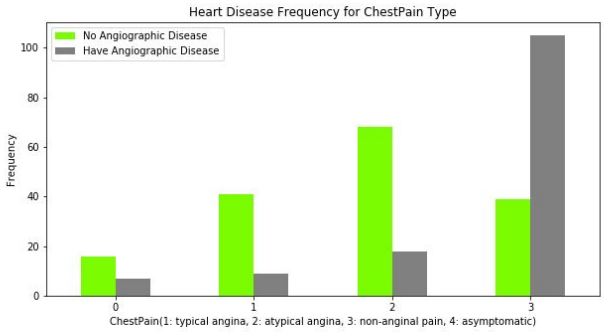
Age	integer
Sex	categorical
ChestPain	categorical
RestBPS	integer
Cholestrol	integer
FBS	categorical
MaxHR	integer
RestECG	categorical
Exang	categorical
OldPeak	float
Slope	categorical
Nmajvess	categorical
Thal	categorical
ADS	categorical

The dataset has only 303 instances of the record with 14 attributes, 13 being the independent variable. The dependent variable had a multiclass distribution which was converted to a binary classified data due to the limitations of data provided. Results achieved for binary class were superior than that were achieved over five classes of dependent variable [11].

The characterization of the Cleveland dataset [13] is given in Table 1.1, it represents the type of data(continuous int/float or categorical) and the brief information of attributes:

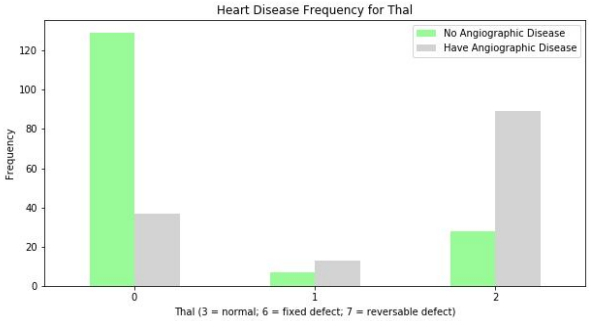
The missing values in some of the research were dropped, making the dataset shorter, with 297 instances instead of 303. Here, the missing values were replaced by nearest of the mean value of categorical variables. Two of Thalassemia values were missing which are replaced by 6 ie fixed defect, and four Namjvess missing values are replaced by 2 ie three major vessels are coloured in fluoroscopy test. In the dependent variable, where narrowing in any major vessel was greater than 50% are replaced by 1.

3.2 Exploration using graphs:



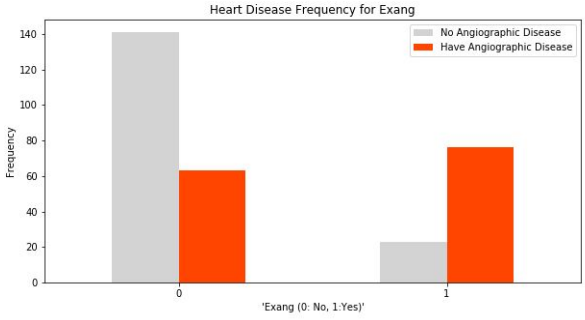
(Figure 1.1: Chest Pain Type vs ADS)

- Figure 1.1 shows that the patients who had asymptomatic type of pain had higher chances of Artery Disease.



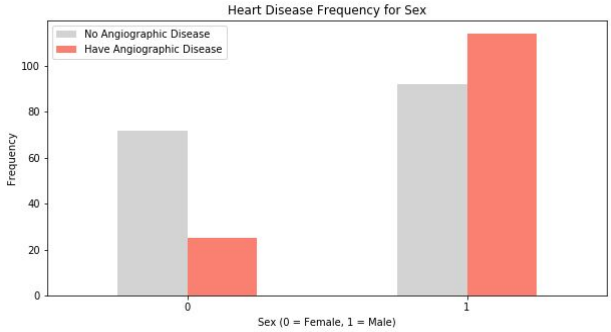
(Figure 1.2: Thalassemia Frequency vs ADS)

- Figure 1.2 depicts that patients that showed normal Thalassemia had lesser chances of Artery Disease.



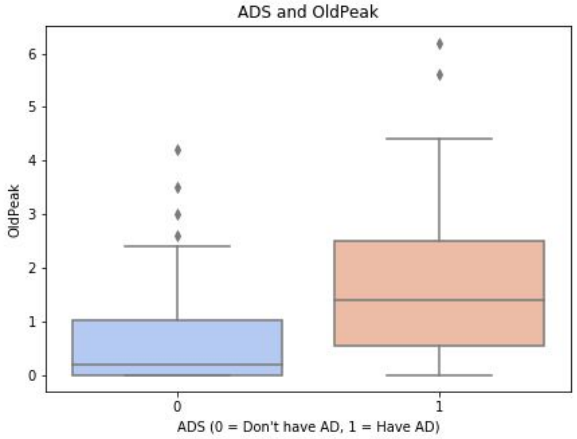
(Figure 1.3: Exercise-Induced Angina Frequency vs ADS)

- Figure 1.3 shows that patients who had exercise-induced angina have higher chances of Artery Disease.



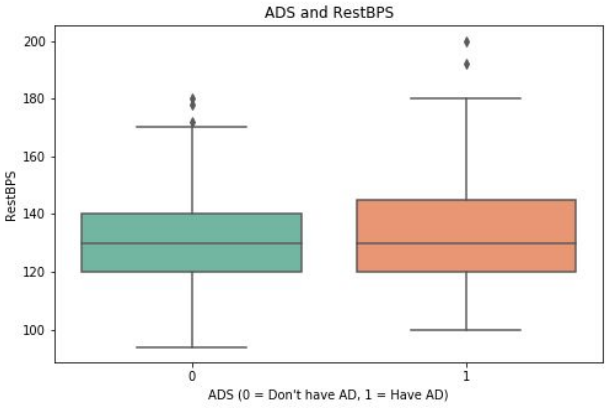
(Figure 1.4: Sex vs ADS)

- Figure 1.4 shows that male patients have significantly higher chances of having Artery Disease.



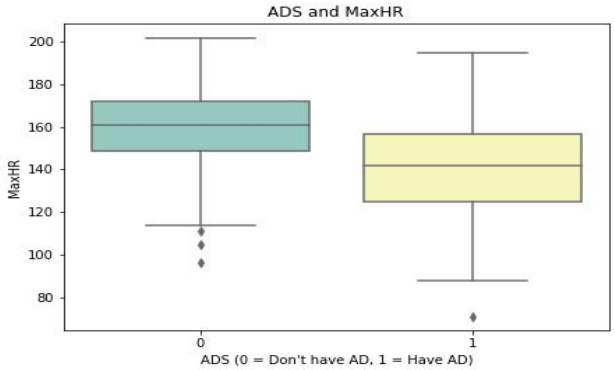
(Figure 1.5: OldPeak and ADS)

- Figure 1.5 shows patients that had higher ST depressions had higher chances of Artery Disease.



(Figure 1.6 RestBPS and ADS)

- Figure 1.6 shows Resting Blood Pressure doesn't have any correlation with Artery Disease.



(Figure 1.7: ADS and MaxHR)

- Figure 1.7 shows patients that achieved higher max. heart rates have lower chances of Artery Disease.

3.3 Data Preprocessing

The train to test ratio is taken as 3:1, and the model was also checked at a 3:2 ratio of train to test. This yielded that there were no significant changes in the performance of models yet, 3:1 train test splitting had way better results than the later.

Before, fitting the model to training set encoding of multiclass categorical variables was done, and all categorical attributes were standardized, using the sklearn library.

4. Methodology:

4.1 Logistic Regression:

Logistic Regression is the simplest way to handle classification problems. In logistic regression a sigmoid function is used instead of a linear function.The value of sigmoid function varies between 0 and 1. It can be used for classification problem as when the value is greater than 0.5 it gives a label 1, otherwise 0.Logistic regression gives a linear classifier boundary.In this experiment logistic regression gave an accuracy of 85.53% on cleveland heart disease dataset.

4.2 K-Nearest Neighbours:

K-Nearest neighbor algorithm is based on instance learning. Instance learning is also called lazy learning as the instances are not processed, they are just stored instead. In this algorithm, a model is not learnt before instead when the test example is provided,it uses the stored examples to classify the class of the test example. This algorithm is based on distance metric. It finds the nearest neighbors depending on the value of k and looks at the value to predict the class of test example.The value of k was set to 5 for experiment and the accuracy obtained on the cleveland heart disease dataset was 88.16%. This also outperformed the other machine learning classification algorithms used.

4.4 Support Vector Machine:

Support Vector Machine(SVM) algorithm is most effective for classification problems.It can be used for linear as well as non-linear classifications depending on the different kernel functions.SVM is an alternative

to bayesian learning. It depends on the support vectors chosen based on their distances.. The points which are close to the decision boundary are known as support vectors. The greater the distance of the point from the margin, the higher is the confidence This model completely depends upon the support vectors chosen and the distance metric.In this experiment, SVM algorithm was applied using default parameters and an accuracy of 88.16% was obtained. This algorithm outperformed all the other classification algorithms.

4.3 Naive Bayes:

Naive bayes algorithm is based on the application of bayes law of probability..Bayes theorem is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

[19]

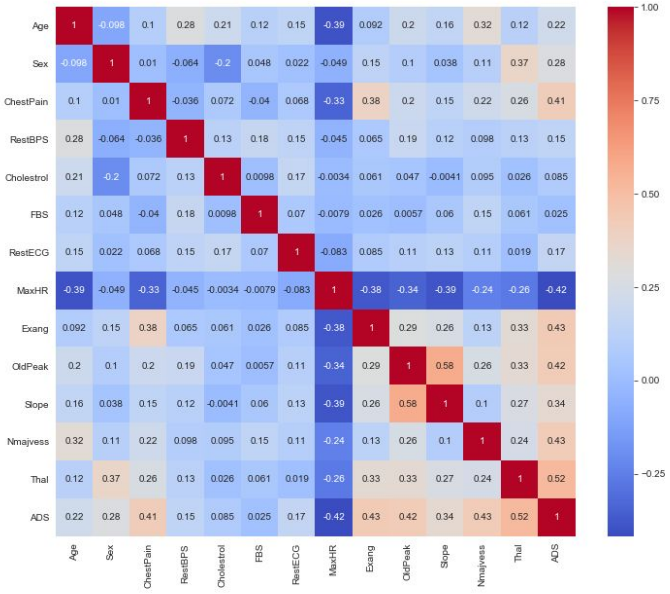
Where A and B are two events. When bayes theorem is applied to classification problems, joint probability has to be calculated. It is very difficult to learn and represent joint probability as there are 2 to the power n possible combinations even if the features are boolean. This makes the calculation and storage intractable. So to simplify this, an assumption is made in naive bayes algorithm. The assumption is that all the features are independent to each other when the class of features are given.Conditional independence is assumed between attributes. This makes Naive bayes a very simple classification algorithm.For n features, only the probability of n-1 features needs to be calculated which can be computed easily.
In this experiment, using the naive bayes algorithm on cleveland heart disease database, accuracy of 84.21% was obtained.

4.5 Decision Tree & Random Forest Classifier:

A decision tree is a classifier in the form of a tree which has two types of nodes, decision nodes and leaf nodes. The decision nodes specify a choice or test. The decision tree is just like any binary tree and can be easily followed to reach a leaf node. It can be used to solve problems of classification and regression,both. Generally, decision trees are used for classification problems. A decision tree can be constructed using two criteria.It can be based on entropy criteria or gini index. A split is made at the internal node if the gini index is low or the information gain is high. External nodes are also known as leafs. External node contains the value or the label also known as the target value. Random Forest uses an army of trees and hence helps in avoiding overfitting in the tree [18, 28]. In Decision tree and random forest parameter 'criterion' was set to entropy , parameter 'max depth of tree' was set to 5 and parameter 'minimum samples leaf' was set to 8. In Random Forest algorithm ,the combination of number of decision trees used were 800(parameter number of estimators).The accuracies obtained were 82.89% and 86.84% on the dataset taken.

4.6 Feature Extraction:

Filter method based on Pearson's Correlation and Backward Elimination Method were used to select the major predicting feature [26]. Using the results of both methods, 8 attributes come out to be major predictors or say most correlated to the dependent variables.
Since, these methods are based on linear relations and continuous variables, therefore, only Naive-Bayes model was able to perform as per expectations.



(Figure 1.8: Heatmap for correlation coefficients)

Relevant Features: ChestPain, MaxHR, Exang, OldPeak, Slope, Nmajvess, Thal, Sex are the union of result achieved from both the methods.

5. RESULTS AND DISCUSSIONS

To assess the performance of various classification models, accuracy, recall and precision were calculated. A comparison was also done by taking 25% and 40% as test data.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$ [14]

Precision and recall are as:

$$\text{Precision} = \frac{tp}{tp + fp}$$ [14]

$$\text{Recall} = \frac{tp}{tp + fn}$$ [14]

5.1 Comparison of models:

Table-5.1 compares the accuracies obtained by the algorithms. These are accuracy obtained before feature selection was applied. It can be seen that KNN and SVM outperforms all other classification models.
Table-5.2 shows the precision and accuracies obtained by different models. It also tells about the number of instances which were classified as true(class-1) and false(class-0).

(Table 5.1: Comparison of different algorithms before optimization(25%))

MODEL	ACCURACY
Logistic Regression	85.53 %
KNN	88.16 %
SVM	88.16 %
Naive-Bayes	84.21 %
Decision Tree	81.58 %
Random Forest	86.84 %

(Table 5.2: Calculated precision and recall of different models(25%))

MODEL	PRECISION	RECALL	PREDICTED FALSE	PREDICTED TRUE
Logistic Regression	85.5	85.5	39	37
KNN	88	88	39	37
SVM	88.5	88	39	37
Naive Bayes	84.5	84	39	37
Decision Tree	81.5	81.5	39	37
Random Forest	87	87	39	37

Table-5.1 and Table-5.2 display the outcomes which were calculated for 25% of test data.Table-5.3 shows the results which were calculated for 40% of test data.

(Table 5.3 Comparison of different algorithms before optimization(40%))

MODEL	ACCURACY
Logistic Regression	82.79 %
KNN	81.97 %
SVM	87.70 %
Naive-Bayes	84.43 %

Decision Tree	79.51 %
Random Forest	85.25 %

(Table 5.4: Comparison of different algorithms after optimization(25%))

MODEL	ACCURACY
SVM	71.05 %
Naive-Bayes	88.16 %
Logistic Regression	89.47 %
KNN	76.32 %
Decision Tree	81.89 %
Random Forest Tree	88.16 %

Table 5.4 shows performance of model after optimization i.e when only 8 major attributes were used for prediction. Only Logistic Regression and Naive-Bayes accuracy was increased significantly by approximately 5% and 4% respectively.

6. CONCLUSION AND FUTURE WORK

Many techniques can be used to predict and prevent coronary heart diseases including machine learning or deep learning techniques.Here six classification algorithms are analyzed to predict CAD: Logistic Regression, KNN, SVM [3], Naive-Bayes, Decision Tree [3], Random Forest. These algorithms are judged on the grounds of outcome accuracy. The data provided was of multiclass classification, due to the small size of data and thirst to achieve better performance, data was converted as binary classified.This study showed that Naive Bayes model turned out to be the best classifier, in accordance with the data, for prediction. This model before feature extraction showed 46 incorrect cases (combining train and test), and only 36 incorrect cases with accuracy of 88.16% thereafter.

There are certainly room for improvement in future. New features can be extracted using feature engineering. Deep learning methods can also be applied to generate new features. PCA can also be used, instead of backward elimination for better results, to reduce the dimensionality of features and figure out informative features. Also, synthetic data, as used by Alfeo Sabay[10], can be a progressive step in disease prediction.

REFERENCES

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Meyer, M., ... & Abi-Mansour, P. (1991). Algorithm to predict triple-vessel/left main coronary artery disease in patients without myocardial infarction. An international cross validation. *Circulation*, *83*(5 Suppl), III89-96.

[2] Alwan, A. (2011). *Global status report on noncommunicable diseases 2010*. World Health Organization.

[3] Kumari, M. and Godara, S. (2011) Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction 1. International Journal of Computer Science and Technology, 2, 304-308.

[4] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, *64*(5), 304-310.

[5] Yao, Z., Liu, P., Lei, L., & Yin, J. (2005, June). R-C4. 5 Decision tree model and its applications to health care dataset. In *Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005.* (Vol. 2, pp. 1099-1103). IEEE.

[6] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.

[7] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications*, *34*(1), 366-374.

[8] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using artificial neural network and feature subset selection. *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, *13*(3).

[9] Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial intelligence*, *40*(1-3), 11-61.

[10] Sabay, A., Harris, L., Bejugama, V., & Jaceldo-Siegl, K. (2018). Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data. *SMU Data Science Review*, *1*(3), 12.

[11] Mehanović, D., Mašetić, Z., & Kečo, D. (2019, May). Prediction of Heart Diseases Using Majority Voting Ensemble Method. In *International Conference on Medical and Biological Engineering* (pp. 491-498). Springer, Cham.

[12] "Heart Disease Data Set, UCI Machine Learning Repository". http://archive.ics. uci.edu/ml/datasets/Heart+Disease"

[13] Heart Disease Data Set of Cleveland, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., PhD.

[14] Wikipedia: https://en.wikipedia.org/wiki/Precision_and_recall#cite_note-OlsonDelen-7

[15] Chen, L., Cao, Q., Li, S., & Ju, X. (2018). Predicting Heart Attacks. International Journal of Computer Applications.

[16] Chaki, D., Das, A., & Zaber, M. I. (2015). A comparison of three discrete methods for classification of heart disease data. *Bangladesh Journal of Scientific and Industrial Research*, *50*(4), 293-296.

[17] Wei, L., & Altman, R. B. (2004). An automated system for generating comparative disease profiles and making diagnoses. *IEEE Transactions on Neural Networks*, *15*, 597.

[18] Sen, S. K. (2017). Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. *International Journal of Engineering And Computer Science*, *6*(6).

[19] Singh, Y. K., Sinha, N., & Singh, S. K. (2016, November). Heart Disease Prediction System Using Random Forest. In International Conference on Advances in Computing and Data Sciences (pp. 613-623). Springer, Singapore.

[20] Basharat, I., Anjum, A. R., Fatima, M., Qamar, U., & Khan, S. A. (2016). A Framework for Classifying Unstructured Data of Cardiac Patients: A Supervised Learning Approach. *framework*, *7*(2).

[21] Hossain, J., FazlidaMohdSani, N., Mustapha, A., & SurianiAffendey, L. (2013). Using feature selection as accuracy benchmarking in clinical data mining. *Journal of Computer Science*, *9*(7), 883.

[22] Chowdhury, D. R., Chatterjee, M., & Samanta, R. K. (2011). An artificial neural network model for neonatal disease diagnosis. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, *2*(3), 96-106.

[23] Chavda, P., Bhavsar, H., Pithadia, Y., & Kotecha, R. (2019). Early Detection of Cardiac Disease Using Machine Learning. *Available at SSRN 3370813*.

[24] Feature Selection with sklearn and Pandas (https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b)

[25] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia Technology*, *10*, 85-94.

[26] Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179-189.

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[28]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[29] Aha, D., & Kibler, D. (1988). Instance-based prediction of heart-disease presence with the Cleveland database. *University of California*, *3*(1), 3-2.

[30] Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial intelligence*, *40*(1-3), 11-61.