Heart Disease Prediction Using Machine Learning Method

Nouh Sabri Elmitwally
Faculty of Computing,
Engineering and the Built
Environment at Birmingham City
University, Birmingham, United
Kingdom.
Nouh.elmitwally@bcu.ac.uk

***Abstract*— The heart disease is also known as coronary artery disease, many hearts affecting symptoms that are very common nowadays and causes death. It is a challenging task to diagnose heart diseases without any intelligent diagnosing system. Many researchers did research on it and developed a diagnostic system to diagnose heart diseases and worked on it. The prediction of cardiovascular disease, required a brief medical history of patients, including genetic information. The world is in acute need of a system for predicting heart disease and it became crucial. Data mining and machine learning are common techniques used in the field of health care to process large and complex data. This research paper presents reasons for heart disease and a model based on Machine learning algorithms for prediction.**

Keywords—coronary, Data mining, cardiovascular, algorithm

I. INTRODUCTION

The heart is the most crucial thing in human existence and its survival. It may be very important for the appropriate functioning of the body. It pumps oxygenated blood to the other components of the body. The heart gets deoxygenated blood and sporting metabolic waste products from the body and sends it to the lungs for oxygenation [1]. If the coronary heart carries out its proper functioning, then it leads to a wholesome existence, however, if the coronary heart forestalls its typical overall performance, it reasons dying of humans. It causes inflammation of blood vessels. Heart diseases cause abnormal blood streaming, which may be very risky for humans [2].

There are many factors that reason coronary heart disease e.g., smoking, excessive blood stress, high cholesterol, diabetes, etc. Smoking is common nowadays in young and aged persons and becoming a trend in youth. Smoking tightens the arteries of the coronary heart and reasons abnormal heartbeat; it additionally increases blood stress. High blood stress reasons many issues it enforces the coronary heart to work harder to deliver blood to the body, it additionally makes the lower-left coronary heart chamber thickens which growth the hazard of coronary heart failure. High blood sugar degrees can harm your nerves or vessels that control your coronary heart.

High cholesterol builds up inside the partitions of arteries, inflicting atherosclerosis, which reasons coronary heart disorder. Due to high cholesterol, arteries turn out to be narrowed, and blood flow is down and blocked. These are the maximum viral illnesses and are very common nowadays and four out of five individuals have those illnesses [3].

In the past, human beings have been unaware approximately the seriousness of coronary heart illnesses and don't be aware of those illnesses, and while some coronary heart illnesses are very deadly to fitness and might even reason be death. Today, most deaths take place because of coronary heart failure. The rate of deaths, because of coronary heart failure, is growing every day. According to the WHO report about 17.9 million people died due to cardiovascular diseases in 2019 [4], which accounts for 32% of all global deaths 85% of these deaths were due to heart attacks and strokes. The reason behind it is that peoples are unaware of the reasons for coronary heart illnesses.

The detection of coronary heart illnesses is turning into very difficult, sluggish, and complex through ordinary clinical strategies in emergent countries because of the loss of professional medical doctors and modernization in exam tools. It is turning into large trouble all around the world [5]. Whereas examination procedures are carried out after studying the medical history of the patient and analyzing symptoms doctors suggest various tests according to the situation of the patient like Blood tests, Electrocardiogram (ECG), coronary angiogram, Exercise stress test, Echocardiogram (Ultrasound), Nuclear Cardiac stress test.

An ECG checks a person's coronary heartbeat, it reads electric impulses of the coronary heart. It's far completed through small sticky dots that are placed-on

palms, legs, and chest then lead connected with an ECG machine which facts' heartbeat which might be withinside the shape of electrical impulses and print them on paper. This is recommended by your doctor while your coronary heart beats abnormally to diagnose coronary heart failure. When you need a clear or detailed image of the heart MRI uses, it is done through magnets and radio waves, it will take moving pictures of the heart on the computer, it is suggested when your physician wants to know how well your heart is working. Coronary Angiogram, this test is taken after a heart attack. A small tube called a catheter is put into an artery on your wrist, arms, and groin and then move inside an artery. In this test, an X-ray of the heart is taken so that a doctor can see blocked arteries [6].

But these strategies have flaws and cannot predict the complexity of the disorder. According to worldwide research in 2019, coronary heart disorder has turned out to be the No. 1 disorder with its diverse variants due to the fact entire body relies upon the right functioning of the coronary heart, if it does not work properly then causes death of a person [7]. The medication can be very costly due to its late diagnosis and lack of medical experts, as late as the disease might be diagnosed, it becomes more and more crucial. The tests carried out nowadays are not able to predict the entire reason for the disorder like in ECG, this isn't always sufficient to diagnose coronary heart disorder more tests are needed for detecting coronary heart problems [8]. A coronary angiogram can influence a patient's kidney [9] it may injure the artery and might reason allergies etc.

The clinical strategies are crucial for diagnosis the coronary heart disorder. The agenda of this paper is to test whether an affected person has a coronary heart disorder or is now no longer with correct strategies. Machine learning is an approach [10]. This is used in the improvement of a computer system wherein algorithms are used, and statistical fashions are utilized to investigate facts without following explicit instructions. Applications of machine learning used in biological databases and increasing day by day. It is helping us in various medical departments and building models that associate a wide range of variables with a disease. SVM can work on small as well as complex datasets and is much more powerful and can be stronger for building machine learning algorithms while the Confusion matrix is used for summarizing the performance of a classification algorithm that your classification model getting right or what kind of error it is making.

In this research paper, machine learning strategies and algorithms of category SVM are used for the detection of coronary heart disorder. For finding the accuracy of data, a data set is selected from Kaggle with patient history and attributes and then preprocess that data. By using these data sets and applying classification algorithms, it can be predicted whether a person has heart disease or not. Data mining is a process of finding exceptions, figures, and associations with large datasets to predict results.

Data preprocessing is the data mining technique that transforms raw data into useful data. In the data preprocessing process first step is data cleaning data may have missing parts so to fill that missing part by the most probable value and remove noisy data by clustering, regression, and binning method in this way first step is done. Data transformation is the next step which transforms the data into suitable form normalization and attribute selection is done in this step. Data reduction is the third and last step in data preprocessing, data cube aggregation, attribute subset selection numerosity reduction and dimensionality reduction are the various steps involves in data reduction [11].

This data is trained using SVM classification algorithms, which are used to check the accuracy of classifying a patient as having a risk of heart disease or not. This method is both cost and memory efficient. [12].

II. LITERATURE REVIEW

Numbers of studies have been done that focus on the diagnosis of heart disease because most of the deaths occurred due to heart failure. Researchers applied different techniques for the diagnosis of heart diseases and achieved different probabilities for different methods. The researchers used data mining methods and algorithms of classification [13]. Decision Tree, Native Bayes, and Neural Network for the prediction of heart diseases. In an experiment, the researcher produced a model using neural network and hybrid intelligent techniques on a dataset which shows a result that the hybrid intelligent techniques give the best result, and the accuracy of prediction is improved by this technique. Machine Learning techniques predict risk at an early stage and are very useful techniques in some techniques.

Researchers used machine learning techniques for the prediction of heart disease some techniques are SVM support vector machine, naive Bayes, neural network, decision tree, and regression classifiers. The researcher shows that SVM [14] is the best technique which gives an accuracy of about 92.1% while neural networks give an accuracy of about 91% and decision trees show lesser accuracy which is about 89.6%. The researchers used the backpropagation algorithm [15] as the best classification technique for the prediction of heart disease. They also proposed a genetic algorithm optimizer against the backpropagation algorithm and the drawback of this is stuck in local minima. They proposed that this methodology give 100% accuracy in the future with fewer errors.

Research is done by S. Prakash et al. in 2017 on heart disease prediction in which they compare two methods Optimality Criterion Feature Selection (OCFS) and rough set feature selection on information entropy (RSFS-IE). The researchers used different types of datasets in terms of computational time, prediction quality, and error rate. They proposed that OCFS is best compared to RSFS-IF because it can take less execution time [16].

Another study was done by researchers in which they took a sample database of the patient record. They trained and tested Neural Network by 13 attributes such as age, blood pressure, angiography report, etc. They recommended supervised network diagnosis of heart

disease and used a backpropagation algorithm for training [17]. The system identified unknown data and set comparisons between unknown data and trained data whenever the doctor fed unknown data and produced a list of probable diseases a patient endangered. The desired output is closest to 100%.

Kim and Kang used a neural network [18] to develop a system for the diagnosis of heart disease. For the detection of features that were more important, sensitivity analysis of features was used. Those features which have high sensitivity were more important than those features which have low sensitivity. By analyzing the change in the sensitivity of features with a change in the value of one feature, correlated features were found after the selection of irrelevant features. Two features were correlated if a change in the value of one feature also change the sensitivity of another feature more than the average change.

K. Polara j et al. compare different algorithms models for the prediction of heart diseases [19] and the result was that multiple linear Regression is better for the prediction of the risk of cardiovascular disease. The study is done by using a dataset that consists of 1000 values. Data was divided into two phases in which 70% of data was used for training of the machine and the remaining 30% was used for the testing purpose, after seeing the result it was confirmed that the regression algorithm is maximum as compared to other models. In 2020 MafizurRehman [20] used the Random Forest algorithm for the prediction of heart disease more effectively and the accuracy was about 97%.

In recent years, for the improvement in the process of prediction of heart diseases, the various system has been proposed using wearable sensors. A wearable medical device-based system was presented by Al-Makhadmeh and Tolba that system collects detail about cardiac patients before and after heart failure. For the appropriate class and valuable function extraction, they carried out function extraction techniques and deep learning models, and then transmitted the collected information to the healthcare system.

That system has limitations because it used 23 attributes for training, and it reasons system complexity and dimensionality. That tool has become no longer accurate, because of inefficient feature extraction and feature weighting approach. A tool that automatically treats coronary heart patients using the net of things (IoT) and a deep learning version called Health fog was presented, and the aim have become to routinely address cardiac affected man or woman information coming from IoT devices [21] Researchers used different algorithms Naïve Byes, Neural Network, Decision Tree, and genetic algorithm for the prediction of heart disease Naïve Bayes shows good results and accuracy was about 96.6% [22]-[23].

III. LIMITATION OF PREVIOUS WORK AND OUR CONTRIBUTION

In previous research, researchers did a great job for the prediction of heart disease using different techniques. Table 1 shows the previous work of researchers.

TABLE I. Analysis of previous Methods and results

Approach	Year	Method used	Results
Vincy Cherian et al,[23]	2017	Naïve Bayes	86%
Rani et al,[18]	2021	Logistic Regression,	86.60%

But those researchers have a few deficiencies. The deficiency of those researchers conquers with the aid of using the methodologies utilized in this article. SVM is utilized in this article to get the preferred result for the prediction of coronary heart disorder. SVM is a computer algorithm that still was efficaciously implemented on a more and more huge variety of organic applications. Support Vector Machine utilized in one-of-a-kind classification problems together with bioinformatics, as an effective machine learning method [24]. Furthermore, this article represents a machine learning model to predict heart disease.

IV. METHODOLOGY

The strategy utilized in this article for the prediction of coronary heart disease is SVM. In the medical field, it is currently a very active research area and in the future, it will be widely used in the biomedical system. SVM is a set of supervised learning, works on small as well as complex datasets. The benefits of support vector machines are that it is effective in excessive dimensional spaces. Mainly due to this the answer isn't restricted to linearity. It is perfect for figuring out sicknesses and the usage of community scans. Since no unique set of rules is needed on how the ailment is diagnosed [25]. The Federated Learning approach is differing from traditional machine learning techniques it's miles a rising approach that could be very useful in price saving and security. In this approach, all datasets submit to a single server [26]. It involves the training of data through several decentralized edges and servers that carry local data samples without sharing them.

A. Preprocessing

Data preprocessing is the step one which initiates the technique. A suitable dataset from Kaggle was taken for this process. In ML it refers to the cleaning and transformation of raw data which is not understandable into readable data and makes it suitable for training machine learning models. Data cleansing, data transformation, and data reduction take place in the technique of data preprocessing. Data cleaning is the process of filling missing values, smoothing noisy data, and removal of outlier's data transformation includes normalization and aggregation while the data reduction process reduced the amount of data, but the result remains the same. The execution of preprocessing necessitates the use of several Kaggle datasets. During this process, data is erased, and lost values and dots are removed to prevent any

inefficiencies and to get higher accuracy. The accuracy depends on given dataset.

B. Dataset Description

A dataset selected from different datasets that are taken from Kaggle has 919 rows and 12 columns which incorporate age, sex, chest ache type, resting BP, cholesterol, fasting BS, Resting ECG, max HR, exercising angina, vintage peak, ST_slop, coronary heart disorder. The coronary heart disorder column has values "1" which suggests the affected person has a coronary heart disorder or "0" which suggests the affected person does now no longer have a coronary heart disorder. The data taken from dataset remains imbalanced, so preprocessing carried out on it. Table 2 below describe the dataset.

TABLE II. Attributes & discerption of heart disease Data set

Feature Name	Type	Detail
Sex	M: male, F: female	Sex of the patient
Age	Integer	Shows age of the patient
Chesting pain type	TA, ATA, NAP, ASY	Patient having which type of pain
Resting BP	Integer	Patient blood pressure level
Cholesterol	Integer	Patient cholesterol level
Fasting BS	0 or 1	Patient fasting blood sugar level
Resting ECG	Either normal or ST	Patient Resting electrodiogram result
Max. HR	Integer	Patient's maximum heart rate
Exercise angina	Y: yes, or N: no	Patient induced angina
Old peak	Numeric value	Patient ST value (measured in depression)
ST_Slop	Up, Flat, down	Slope of peak exercise
Heart Disease	0 or 1	Heart disease or not

C. Proposed Model

Proposed Model that is applied in this article is using the method of SVM to get better accuracy. As compared with the last implementations and studies, the models considered for implementation in this article give a better-optimized result.

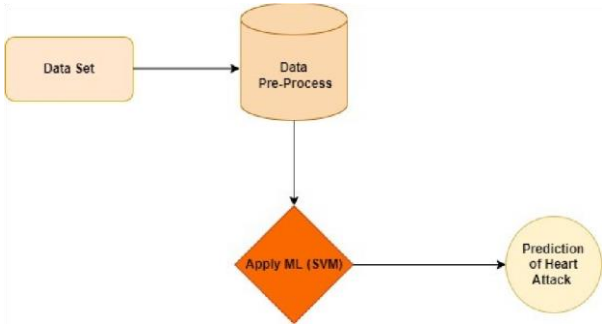

Fig. 1 Proposed Model for Heart Disease Prediction

In this Model as shown in figure1, by applying first the preprocessing on data set and label it with different form then applying the SVM trained model on it for better accuracy and then analysis these results with using Confusion Matrix for verifying. In this Model, by applying first the preprocessing on given data set and label it with a different form then applying the SVM trained model on it for better accuracy, and then analyzing these results using Confusion Matrix for verifying.

D. Implementation Detail

All Results were carried out on a computer with a Core i6 processor, 16 GB RAM, and Super GPU at 3.60 GHz. MATLAB 2020a is used for this purpose and implementation.

V. STIMULATIONS AND RESULTS

The method and implementation implemented here is SVM. Support Vector Machines (SVMs) are a set of supervised learning methods for classification, regression, and outsourced discovery. The advantages of support vector machines are valid in high-dimensional spaces. It still works when the number of dimensions exceeds the number of samples. Figure 2 and Table 3 below show the results of applying SVM on dataset.

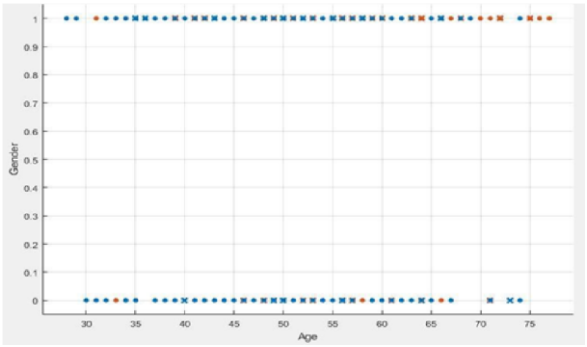

Fig.2 Scotter Plot Diagram of SVM Trained Model

TABLE III. Accuracy obtained using SVM

Training Time	Training Accuracy	Test Time	Test Accuracy
30.533 Sec	90.5%	6.932 Sec	78.7%

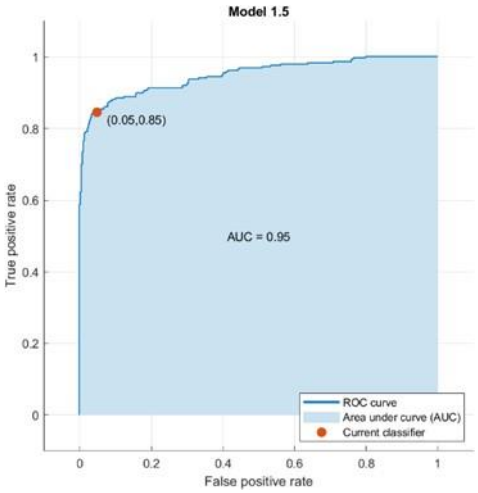

Fig. 3 ROC Curve of Proposed SVM Trained Model

Figure 3 shown the performance of the SVM algorithm. The AUC- ROC (Area Under the Curve - Receiver Operating Characteristic) curve is a performance measure for classification tasks at various thresholds. ROC is the base curve and AUC stands for class score or scale. It shows the ability of the model to distinguish between categories. The model better represents 0 as 0 for AUC and 1 as 1. In terms of measurements, the model is better than AUC in distinguishing patients with AUC from patients without any disease.

A. CONFUSION MATRIX:

The confusion matrix provides a comprehensive overview of model performance. Contrary to its name, you will realize that the confusion matrix is a simple yet powerful concept. The confusion matrix is an N x N matrix used to evaluate the performance of a classification model, where N is the number of target groups. This matrix compares the actual target value with the predictions made by the machine learning model. Table 4 shows the results of Confusion Matrix.

$$Accuracy = \frac{(TP+TN)}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$PPV = \frac{TP}{TP+FP} \quad (2)$$

$$TNR = \frac{TP}{TP+FP} \quad (3)$$

$$TPR = \frac{TP}{TP+FN} \quad (4)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

$$F1\ Score = 2 \times \frac{PPV \times TPR}{PPV \times TPR} \quad (6)$$

TABLE IV. Performance of SVM Confusion Matrix

Sensitivity / TPR	93.44%
Specificity / TNR	88.45%
Precision / PPV	84.62%
Accuracy	90.47%
F1 Score	88.81%
Matthews Correlation Coefficient	80.84%

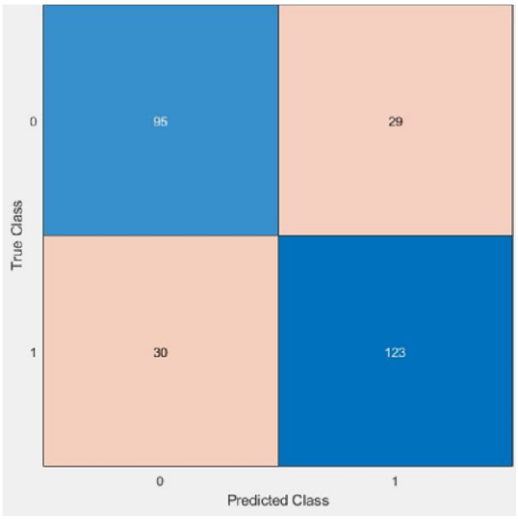

Fig. 4 Confusion matrix of SVM Model

Figure 4 shows Confusion Matrix precision indicates how many of the instances that were correctly predicted turned out to be positive. Recall indicates how many of actual positive cases this model was able to properly predict. In essence, when by seek to enhance model's precision, the recall suffers, and vice versa. In a single value, the F1-score includes both trends.

VI. CONCLUSION

Many people suffered heart damage as a result of major coronavirus epidemic, according to experiments. As a result, research is warranted to develop a suitable diagnostic method that focuses on the incidence of heart failure and can detect it early enough to prevent death. It assists patients in diagnosing heart illness regarding medical information from past heart disease diagnoses. The SVM approach was used to build this model. The model has a 90.47 percent accuracy. Using additional training data raises the risk of the model correctly detecting cardiac illness. To simplify data and compare outcomes, several methods might be performed. Additional techniques to link trained ML and DL cardiac models with specific multimedia can be found for the convenience of patients and clinicians.

REFERENCES

[1] Heart and Circulatory System (for teens) - Nemours. Kidshealth.org. Https://kidshealth.org/Nemours/en/teens/heart.html [accessed 2 December 2021].

[2] Sahoo, P. & Jeripothula, P. (2020) Heart Failure Prediction Using Machine Learning Techniques. SSRN Electronic Journal.

[3] Javeed, A., Rizvi, S., Zhou, S., Riaz, R., Khan, S. & Kwon, S. (2020) Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification. Mobile Information Systems 2020, 1-11.

[4] Chicco, D. & Jurman, G. (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision

[5] Anon. (2021). Https://www.healthywa.wa.gov.au/Articles/A_E/Commonmedical-tests-to-diagnose-heart-conditions [accessed 2 December 2021].

[6] Shu, T., Zhang, B. & Tang, Y. (2017) Effective Heart Disease Detection Based on Quantitative Computerized Traditional Chinese Medicine Using Representation Based Classifiers. Evidence-Based Complementary and Alternative Medicine 2017, 1-10.

[7] Anon. (2021) The top 10 causes of death. Who.int. Https://www.who.int/news-room/fact-sheets/detail/the-top-10causes-of-death [accessed 2 December 2021].

[8] Risks. Stanfordhealthcare.org.Https://stanfordhealthcare.org/medical tests/e/ekg/risks.html [accessed 2 December 2021].

[9] Coronary angiogram - Mayo Clinic. Mayoclinic.org. Https://www.mayoclinic.org/tests-procedures/coronaryangiogram/about/pac-20384904 [accessed 2 December 2021].

[10] MachinelearningWikipedia. En.wikipedia.org.Https://en.wikipedia.org/wiki/Machine_learning [accessed 2 December 2021].

[11] Data Preprocessing in Data Mining - GeeksforGeeks. GeeksforGeeks. Https://www.geeksforgeeks.org/datapreprocessing-in-data-mining/ [accessed 2 December 2021].

[12] Goel, R. (2021) Heart Disease Prediction Using Various Algorithms of Machine Learning. SSRN Electronic Journal.

[13] Anon. (2022) WCECS2014 pp809-. Academia.edu. Https://www.academia.edu/35720965/WCECS2014_pp809_ [accessed 1 January 2022].

[14] Latah, C. & Jeeva, S. (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 16, 100203.

[15] Nanekar, G. (2021) Heart Disease Prediction using Neural Network. International Journal for Research in Applied Science and Engineering Technology 9, 1907-1910.

[16] Anon. (2022) Improving Heart Disease Prediction Using Feature Selection Approaches. Ieeexplore.ieee.org. Https://ieeexplore.ieee.org/abstract/document/8667106/ [accessed 1 January 2022].

[17] Gavhane, A., Kokkula, G., Pandya, I. and Devadkar, K., 2018, March. Prediction of heart disease using machine learning. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1275-1278). IEEE.

[18] Rani, P., Kumar, R., Ahmed, N. & Jain, A. (2021) A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments 7, 263-275.

[19] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P. & kumar, N. (2021) Latest trends on heart disease prediction using machine learning and image fusion. Materials Today: Proceedings 37, 3213-3218.

[20] Pavithra M., M. (2022) Effective Heart Disease Prediction Systems Using Data Mining Techniques. Annalsofrscb.ro. Https://www.annalsofrscb.ro/index.php/journal/article/view/2172 [accessed 2 January 2022].

[21] Ali, F., El-Sappagh, S., Islam, S., Kwak, D., Ali, A., Imran, M. & Kwak, K. (2020) A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion 63, 208-222.

[22] Https://www.researchgate.net/profile/Anbarasi Masilamani/publication/50361284_Enhanced_Prediction_of_Heart _Disease_with_Feature_Subset_Selection_using_Genetic_Algorith m/links/54accada0cf2479c2ee853b1/Enhanced-Prediction-ofHeart-Disease-with-Feature-Subset-Selection-using-GeneticAlgorithm.pdf [accessed 6 January 2022].

[23] Anon. (2022). Ijcstjournal.org. Http://www.ijcstjournal.org/volume-

[24] 5/issue-2/IJCST-V5I2P13.pdf [accessed 6 January 2022].

[25] Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology, 24(12), pp.1565-1567.

[26] Matveeva, N. (2021) ARTIFICIAL NEURAL NETWORKS IN MEDICAL DIAGNOSIS. System technologies 2, 33-41.

[27] Anon. (2022) Analysis of Neural Networks Based Heart Disease Prediction System. Ieeexplore.ieee.org. Https://ieeexplore.ieee.org/abstract/document/8431153/ [accessed 9 January 2022].