Heart Disease Prediction using Machine Learning

Mrinal Anand
BTech CSE
DIT University, Dehradun
1000012534@dit.edu.in

Nikhil Anand
BTech CSE
DIT University, Dehradun
1000013787@dit.edu.in

Hari Om Gupta
BTech CSE
DIT University, Dehradun
1000013143@dit.edu.in

Author: - Ms. Kritika Joshi
Department of Computer Science & Engineering
DIT University, Dehradun

***Abstract*— Since there are more and more occurrences of heart disease every day, it is important and concerning to anticipate any prospective problems. Heart disorders, which are roughly categorized as a variety of abnormal heart problems, are thought to be the cause of about 1 in every 4 fatalities. Following the brain, which is the most important organ in the human body, comes the heart. It circulates the blood and feeds it to all of the body's organs. It is known that 32 % of fatalities occur due to heart condition and they are also the main factor in a substantial number of deaths—more than 17.9 million in 2019—as well. Therefore, a system that can predict with exact precision and dependability is required for the appropriate and prompt diagnosis as well as the treatment of such diseases. Heart condition diagnosis is a difficult process that demands both quickness and precision. This study focuses on focusing people who are more prone to develop heart condition based on a number of medical characteristics (age, sex, chest pain type, resting blood pressure, cholesterol, heart rate etc). The objective of this study is to create a trustworthy method for predicting heart failure patients using their clinical records and laboratory data. The suggested work uses various techniques such as Decision Tree, Logistic Regression, Support vector machine, Random Forest, K-Nearest Neighbor and Gaussian NB algorithm to forecast the likelihood of Heart Disease and classify patients risk levels. A very beneficial technique was used to control how the model can be used to increase the accuracy of prognosis of Heart Disease in any individual. This paper compares the performance of various machine learning algorithms.**

Keywords: Heart disease; Machine learning; Random Forest; Logistic regression; Support Vector Machine; Decision Tree; K-nearest Neighbor; Gaussian NB; SVC; CVD

I. INTRODUCTION

Machine learning is a very diverse and wide-ranging field, and its use and scope are expanding every-day. To forecast the future, machine learning employs a range of classifiers from supervised and unsupervised learning to prognosis the outcomes and examine the accuracy of a dataset. There are many sectors in the healthcare industry where ML has shown to be highly helpful. Given the exponential rise in the amount of real-time digital data produced by the medical industry (such as through Health Records Devices, wearable technology, diagnostic report, etc.) [11], to process such data, intelligent Systems must be created. Such information can be used for a range of studies, including ones on population ageing, the success of newly developed treatment regimens, age-related health practises, and medical spending reports. In the long run, these analyses would give information to help organisations, individuals, and governments in the area create improved medical policies and improve the ones already in place[12]. Low socioeconomic status individuals do not receive the best care for cardiovascular disease, which results in significant fatalities and the identification of individuals at risk for cardiovascular disease. Recent research indicates that the leading cause and disease of fatalities in people worldwide is cardiovascular disease[13]. According to the WHO, there will be a sharp rise in mortality related to heart disease, and various factors have been identified as being harmful to the heart [14]. According to World Health Organization (WHO) estimates, heart-related and cardiovascular disorders cost India $237 billion between 2005 and 2015. Therefore, it is crucial to have a realistic and accurate prognosis for heart-related diseases[14]. How effectively a condition will be addressed overall depends on when it is discovered. To avoid bad or tragic results, the proposed approach tries to detect certain cardiac problems early. There are records of a significant collection of medical information gathered by medical experts that can be analysed and mined for insightful information. The use of data mining techniques enables the extraction of crucial and concealed information from the enormous amount of accessible data. Large, well-formatted datasets are easily handled by ML, a branch of data mining. A range of illnesses may be identified, detected, and predicted using machine learning in the medical sector [11]. Worldwide, medical groups gather information on a range of health-related topics. To obtain insightful knowledge from these data, here are several machine learning methods available. However, the amount of data gathered is enormous, and it is frequently very chaotic. These datasets can be readily explored using a variety of machine learning methods, even though they are too large for human brains to understand. In order to learn about heart condition at an early stage, this research analyses the performance of several ML approaches, including DT, LR, RF, SVM, KNN and Gaussian NB and also these methods are cost-efficient.

II. LITERATURE REVIEW

Since a decade ago, the study of healthcare has received great attention. In the healthcare sector, almost all algorithms are employed and effectively tested. Several researchers have analysed ML techniques as a tool to improve survival prognosis in patients. The majority of these researcher mainly focused at identifying the primary risk factor that mortality in

heart failure patients. In [1], researcher Vardhan Shorewala used a huge dataset (70000+ variables) to develop six machine learning models, which were then compared for accuracy, specificity and sensitivity for the optimal use. Waigi, R[2] predict the accuracy with decision tree which results in low accuracy score although Waigi, R mainly focused on a device for continuous monitoring while using the dataset which will improve the accuracy. Seyedamin Pouriyeh[5] implemented the seven classifier although four were same as other but the unique ones were using ensemble learning techniques including bagging, boosting, and stacking, and merging the Cleveland dataset's radial basis function, a single conjunctive rule learner, and multilayer perceptron models. Single Conjunctive Rule Learner had the lowest accuracy (69.96%) and Support Vector Machine had the best accuracy (84.15%), after applying bagging Support Vector Machine got the same and Decision Tree got the worst with 78.54%. In [15], a genetic algorithm that had been taught using a neural network was used to detect cardiac illness with an accuracy of 97.8%. In the paper[16], three distinct classifiers, namely KNN, Decision trees, and NB classifiers, were used to divide the data into heart disease risk and non-risk categories. The mortality rate from heart disease has decreased as a result of the usage of fuzzy logic and rough set techniques in [17][18].

Hasan and Bao (2020) [19] conducted a study with the primary goal of discovering the best effective feature selection strategy for forecasting cardiovascular illness through a comparison of several algorithms. First, the filter, wrapper, and embedding three well-known feature selection techniques were considered. Then, using a common "True" criterion based on a Boolean method, a feature subset was obtained from these three techniques. This method required two steps of feature subgroup retrieval. To support the comparison precision and determine the best predictive analytics, a variety of models were taken into account, including XGBoost, KNN, RF, and SVC. To compare all properties, the artificial neural network (ANN) was employed as a benchmark. According to the data, the wrapper technique combined with the XGBoost classifier produces the best precise forecast findings for cardiovascular illness. Accuracy was provided XGboost at 73.14%, Artificial Neural Network at 73.20% and SVC at 73.18%.

The aim of research by Drod et al. (2022) [20] was aimed to identify the most important risk factors for CVD in people with metabolic-associated fatty liver disease using machine learning approaches. Blood biochemistry was investigated for 191 individuals with metabolic-associated fatty liver disease, and asymptomatic atherosclerosis was assessed. ML techniques, including principal component analysis, univariate feature ordering, and multiple logistic regression classifier, were used to create a model to identify those who have the highest risk of developing cardiovascular disease. According to the study, the three most important clinical traits were hypercholesterolemia, plaque levels, and length of diabetes. With an AUC of 0.87, the ML method worked well, correctly classifying 114/144 (79.17%) low-risk patients and 40/47 (85.11%) high-risk patients. The study's outcomes imply that a machine learning method is useful for getting individuals with extensive cardiovascular illness who have metabolic-associated fatty liver disease based on simple patient criteria.

Shah et al.'s research from 2020 [21] sought to create a model for forecasting cardiovascular illness using ML methods. The Cleveland heart disease dataset, which was taken from the ML repository, had 303 occurrences and 17 characteristics that were utilised to create the data for this research. A variety of supervised categorization methods, such as KNN, NB, DT,and RF were utilised by the authors. The study's findings showed that, at 90.8% accuracy, the k-nearest neighbor model had the greatest degree of precision. The research emphasises the possible value of machine learning methods for forecasting cardiovascular disease and stresses the significance of picking the right models and methods to get the best outcomes.

III. DATASET

The dataset contains the medical histories of 304 distinct patients, all of varying ages. This information on the patient's medical history, including their age, resting blood pressure, fasting blood sugar level, etc., provided by this collection help us identify patients who have been identified with cardiac disease or not. This collection allows for the extraction of the pattern that identifies patients at risk for developing cardiac disease. These documents are divided into two sections: Testing and Training. The UCI Heart Disease Dataset, an open dataset gathered from the UCI ML Repository, was selected for this study. There are 303 records in the collection. Only 13 characteristics, including one target quality, out of a total of 76 traits, were considered in this study.

Table 1 lists the eight category and six numerical properties of the UCI dataset.

Table 1. Attribute description of the UCI heart disease dataset

Sr no.	Attribute	Description
1	Age(age)	Age of the patient (in years)
2	Sex(sex)	Gender (0= Female and 1=Male)
3	Chest Pain(cp)	1=Typical angina,2=Atypical angina,3:Non-anginal pain,4:Asymtomatic pain
4	Resting Blood Pressure(trestbps)	Resting Blood Pressure(in mm Hg)
5	Serum Cholesterol(Chol)	Serum Cholesterol level (in mg/dl)
6	Fasting Blood Sugar(fbs)	Fasting Blood Sugar(>120mg/dl 0=False , 1-True)
7	Rest Electrocardiograph(restecg)	Resting ECG (0=Normal,1=ST-T wave abnormality, 2=LV Hypertrophy)
8	Maximum Heart Rate(thalach)	Maximum heart rate achieved
9	Exercise-Induced angina(exang)	Exercise-Induced angina (0=No, 1=Yes)
10	ST Depression(oldpeak)	ST depression induced by exercise relative to rest
11	Slope of ST segment(slope)	Slope of peak exercise ST segment(1=up sloping, 2= flat, 3=down sloping)
12	Vessel count(ca)	Number of major vessels colored by fluoroscopy(range 0-3)
13	Thalassemia(thal)	Thalassemia type (normal, fixed defect, reversible defect)
14	Heart Disease(target)	0= negative of disease,1=positive for heart disease

Fig.1- Various Attribute used are listed

IV. METHODOLOGY

This research seeks to estimate the likelihood of developing heart disease using robotic heart disease prognosis, which may be useful for patients and medical professionals. The examination of several methods, including KNN, RF, LR, Gaussian NB, DT, and SVM, is presented in this work and after pre-processing of data, dataset implemented on these classifiers and the results are in the research paper. The data got cleaned up and removed unwanted information in order to improve the accuracy or approach. The proposed model is subsequently put into practice, and its efficacy and accuracy are evaluated using a range of performance criteria. For prognosis, this model makes use of 13 factors, including age, sex, blood pressure, cholesterol, and fasting sugar.

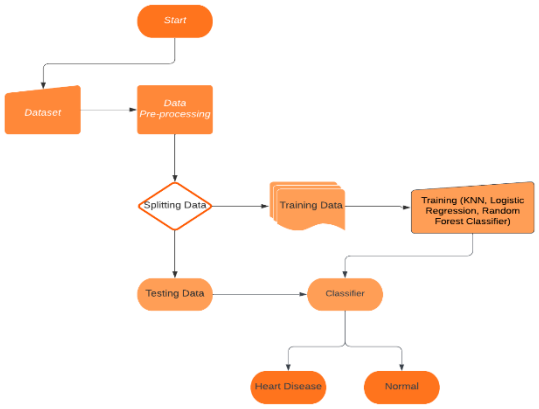

Fig.2- Proposed Model

4.1 Classification Algorithm

An approach to guided learning used to forecast outcomes from historical data is classification. The method for diagnosing cardiac illness using classification methods is suggested in this article. A classifier is taught individually using the training dataset, It was divided into a test set and a training set. On the dataset, various classifiers are used to determine their performance.
In every case the performance is determined by applying the accepted measures such as accuracy, recall, Precision and F-measure which are evaluated using predictive classification table, Confusion Matrix.
In the section that follows, classifier operation is discussed.

4.1.1. Decision Tree

Decision tree is simple to construct and evaluate the data in a graph with a tree-like shape when using a decision tree to create structures that resemble trees. Recursively dividing the dataset into subgroups depending on the values of one or more input variables is how the decision tree method operates. Each split is chosen to maximize the knowledge acquisition, which is a measure of how much the split reduces the uncertainty in the prognosis. The loop continues until a criteria is met, like when all the data in a subset belongs to the same class or when a maximum depth of the tree is reached. Once the decision tree is constructed, for new patients, it can be used to gauge their risk of developing heart disease based on their input data. One potential advantage of using a decision tree for heart condition prognosis is that it can be easier to interpret than other machine learning algorithms, since the tree structure can be represented graphically and the decision-making process is explicit. However, decision trees can also be prone to overfitting and maybe difficult to generalize to new data, so it is important to carefully validate the model and tune its parameters to ensure reliable performance. The alteration in entropy when decision tree nodes are used to break training instances into smaller groups. Information gain is the measurable change in entropy.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

The decision tree managed to reach an accuracy of 77.8%. The decision tree classifier obtained 73.0% accuracy in [1] study and 72.77% accuracy in [2] research in terms of accuracy.

4.1.2. Logistic Regression

A statistical technique frequently when classifying things in binary tasks is logistic regression, where the goal is to predict whether an instance belongs to one of two classes. In the case of heart disease prognosis, the two classes might be "has heart disease" and "does not have heart disease". To build a logistic regression model for heart disease prognosis, one would typically start by collecting a dataset of instances with known heart disease status and a set of relevant features (such as age, gender, blood pressure, cholesterol level, etc.). The dataset would be split into a training and a testing set, and the logistic regression model would be trained on the training set using a suitable optimization algorithm (such as gradient descent). Once trained, the model may be used to predict outcomes using brand-new data. The threshold value is used by the model to produce a binary prognosis and to determine the chance that each occurrence belongs to the positive class.
An accuracy of 85.24% has been achieved by Logistic Regression.

4.1.3. Support Vector Machine

The widely used supervised learning method known as Support Vector Machines (SVMs) is utilized for classification tasks, including heart disease prognosis. The algorithm identifies a hyperplane that maximally separates instances with and without heart disease based on the feature values. Once the SVM is trained, it can be used to make prognosis on new instances. The SVM calculates the distance of the new instance to the hyperplane and assigns it to the positive or negative class based on its distance and the chosen threshold value.
An accuracy of 86.88% has been achieved by Support Vector Machine. In the People's Hospital dataset, Shan Xu et al. employed SVM to reach an accuracy of 98.9% [3]. SVM performs best in [4] with 85.7655% of instances properly identified, while SVM is combined with boosting approach in [5] to get an accuracy of 84.81%.

4.1.4. Random Forest -

Random Forest Classifier [6] are based on decision tree. With the help of the ensemble learning algorithm Random Forest, classification tasks, such as predicting heart disease, may be performed more accurately and robustly by combining several decision trees. On portions of the training data and feature subsets, Random Forest creates numerous decision trees, and then combines the forecasts of the different trees to

provide a final prognosis. This helps to approach enhance the model's generalization capabilities by decreasing overfitting of the model. Each decision tree in the Random Forest algorithm predicts the heart disease status of an instance based on a subset of features. Once the Random Forest is trained, it can be used to make prognosis on new instances by passing them through each decision tree in the forest and aggregating the prognosis using majority voting. It performs effectively with huge datasets of great depth. The findings are ambiguous and the forecasting process is time-consuming since it requires vast data sets and several trees.

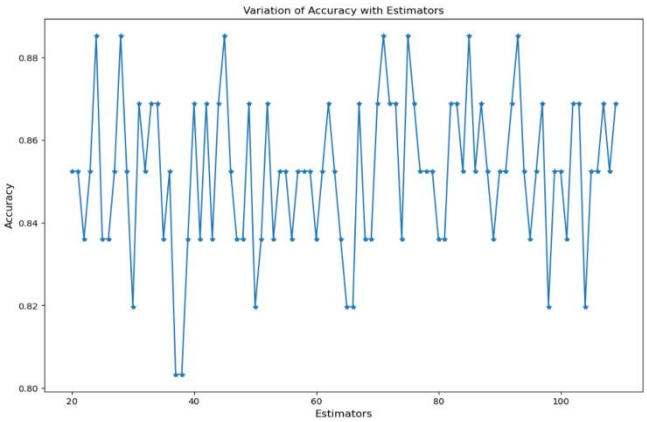

Fig.3- Variation of Accuracy with Estimators

An accuracy of 88.52% has been achieved by Random Forest Classifier. The accuracy of random forest's prognosis of coronary heart disease in [10] is 97.7%.

4.1.5. K-Nearest Neighbor-

The KNN model, one of the most straightforward and extensively used supervised classification algorithms, was proposed by Fix and Hodge. Heart disease can be predicted using the K-Nearest Neighbor (KNN) [7] algorithm, a non-parametric classification technique. KNN may be trained on a dataset of cases with known heart disease status and a set of pertinent attributes in the context of predicting heart disease. (such as age, gender, blood pressure, cholesterol level, etc.). The algorithm determines the separation between each instance in the training set and the new instance. The class of the new instance is allocated based on the majority class among its k-nearest neighbours after choosing the k instances with the smallest distances [8]. KNN utilises a variety of metrics, including Euclidean, Manhattan, and Minkowsky, when working with numerical data[9]. Even KNN is a simple and effective algorithm for heart disease prognosis, but may require careful tuning of the hyperparameters and careful preprocessing of the data.
An accuracy of 93.4% has been achieved by K-Nearest Neighbor.

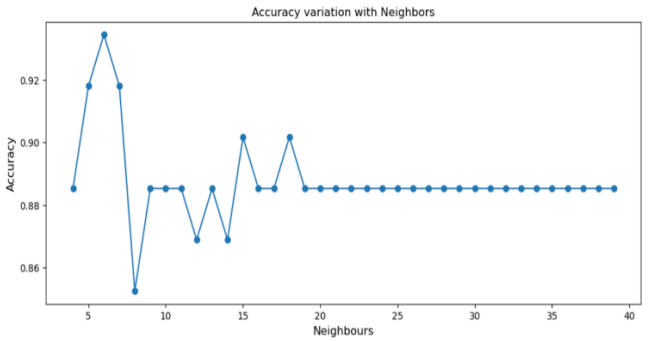

Fig.4- Accuracy Variation with Neighbors

4.1.6. Gaussian NB-

Heart disease can be prognosed using the probabilistic classification process known as Gaussian NB (GNB). Using Bayes' theorem to determine the posterior probability of each class given the feature values of a new instance, the algorithm simulates the probability distribution of each feature for each class. The conditional probability of each feature given each class is calculated by GNB together with the prior probability of each class (i.e., the frequency of each class in the training data). The posterior probability of each class is then determined using Bayes' theorem given the feature values of a new instance. The class with the maximum posterior probability is assigned to the new instance. By computing the PP of each class given the feature values and assigning the class with the greatest probability, the GNB model may be used to make prognosis on new instances once it has been trained. Although, GNB is a simple and effective algorithm for heart disease prognosis but may require careful consideration of the underlying assumptions and potential limitations.
Gaussian NB has reached an accuracy of 86.8%.

V. RESULT

For this study, an Intel Core i5 quad-core processor running at 1.4 GHz with 8 GB of RAM was used. There are 303 entries in the collection. Only 13 qualities—out of a total of 76—include one target trait. This study's objective is to assess the performance of several categorization methods in order to identify the best reliable classification algorithm for predicting whether or not a patient would have cardiac disease. The UCI dataset was used in this study to test a variety of classification algorithms, including DT, LR, RF, SVM, KNN and Gaussian NB. According to the findings, when compared to other categorization algorithms, KNN provides the greatest accuracy.

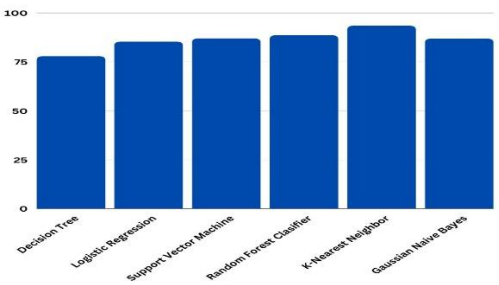

Fig.5- Performance of Classifiers

VI. CONCLUSION

Machine learning is important when it comes to illness prognosis. Various ML methods are used in this study to forecast heart disease on Cleveland dataset. Each of the previously mentioned algorithms has done incredibly well in some situations while failing miserably in others. K-Nearest Neighbor did well while decision tree performs poorly. The fact that the model was built on a relatively smaller dataset due to the absence of an openly accessible dataset for this reason is a drawback of the present research. A sizable sample from a diverse geographic area would undoubtedly increase the model's robustness and give researchers a better grasp of the characteristics that are most likely to lead to mortality in patients with heart failure. By using these above techniques prognosis of patient will become faster, more effectively, and with a significant cost reduction. A machine-learning risk stratification algorithm may enhance therapeutic practice, patient care, and overall results.

Despite the positive outcomes, there are a number of restrictions that should be taken into account. First of all, because only one dataset was used in the research, it may not be applicable to other demographics or patient groups. Additionally, the research only took into consideration a small number of clinical and demographic characteristics and ignored other possible risk associated for heart conditions, such as hereditary predispositions or aspects of lifestyle.

REFERENCES

[1] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655. [Google Scholar] [CrossRef]

[2] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645. [Google Scholar]

[3] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prognosis Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.

[4] Hanen Bouali and Jalel Akaichi et al. “Comparative study of Different classification techniques, heart Diseases use Case.”, 2014 13th International Conference on Machine Learning and Applications.

[5] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “ A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017.

[6] Leo Breiman, Random forests. Machine Learning, 45(1):5–32, 2001.

[7] Huafeng Liu, Chao Li, Shuheng Zhang, Huan Zhang, Lifang Pang, Kinman Lam, Chun Hui, and Su Zhang. Using the k-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. Computational and Mathematical Methods in Medicine, 2012:876545, 2012.

[8] Swain, Debabrata, Santosh Pani, and Debabala Swain. 2019, April. An efficient system for the prognosis of Coronary artery disease using dense neural network with hyper parameter tuning. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 (6S). 689-695.

[9] Ghumbre, S.U. and A.A. Ghatol. 2012. Heart disease diagnosis using machine learning algorithm. In Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) Visakhapatnam India,Springer, 132,217-225.

[10] Ahmad Shahin, Walid Moudani, Fadi Chakik, Mohamad Khalil et al. ”Data Mining in Healthcare Information Systems: Case Studies in Northern Lebanon”, ISBN: 978-1-4799-3166-8 ©2014 IEEE.

[11] K. Shailaja, B. Seetharamulu and M. A. Jabbar. (2018) "Machine Learning in Healthcare: A Review," Second International Conference on Electronics, Communication and Aerospace Technology (ICECA): 910-914.

[12] M. A. Khan and F. Algarni. (2020) "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," IEEE Access 8: 122259-122269.

[13] Centers for Disease Control and Prevention (CDC). Deaths: leading causes. Available: https://www.cdc.gov/nchs/fastats/leading-causes-of-death

[14] World Health Organization(WHO), Cardiovascular Diseases(CVDs) [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[15] KaanUyar Ahmet Ilhan, “Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks” 9th international conference on theory and application of soft computing, computing with words and perception. Budapest, Hungary: ICSCCW; 2017. 24-25 Aug

[16] Y. Alp Aslandoganet. al., “Evidence Combination in Medical Data Mining”, Proceedings of the international conference on Information Technology: Coding and Computing (ITCC’04) 0-7695-2108-8/04©2004 IEEE.

[17] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, “Fuzzy logic based decision support system for component security evaluation,” International Arab Journal of Information Technology, vol. 15, pp. 1–9,2015.

[18] S. Nazir, S. Shahzad, and L. Septem Riza, “Birthmark-based software classification using rough sets,”Arabian Journal for Science and Engineering, vol. 42, no. 2, pp. 859–871, 2017

[19] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prognosis. Health Technol. 2020, 11, 49–62. [Google Scholar] [CrossRef]

[20] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240. [Google Scholar] [CrossRef] [PubMed]

[21] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prognosis using Machine Learning Techniques. SN Comput. Sci. 2020, 1, 345. [Google Scholar] [CrossRef]