Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3, June 2021:993- 1001

Research Article

Cardiovascular heart disease prediction using machine learning classifiers with data mining techniques

[1,*]Syed Shareefunnisa, [2]Devarapalli Retz Mahima, [3]Dr Y Padma

Abstract

Cardiovascular disease is among the common causes of mortality rate in the world. It is tough for healthcare professionals to forecast because it is a complex undertaking that necessitates competence and a greater level of information. The medical system is still characterized by an abundance of information but a scarcity of knowledge. Mostly online, there is a wealth of information about healthcare organizations. However, reliable analytic techniques for uncovering underlying correlations and relationships between variables are lacking. A computerized medical diagnostics system would promote health productivity while lowering costs. Cardiovascular attack detection required a huge amount of information that is very much complicated and enormous to collect and analyzed utilizing traditional existing functions. Our research aim is to detect the most suitable machine learning methodology for finding cardiac disease which is operationally efficient and accurate. We created a heart disease prediction program that utilizes the patient history to forecast whether or not a person would be identified with cardiovascular disease. To identify and categorize patients having cardiovascular disease, we applied various machine learning methods such as logistic regression & K - Nearest Neighbours (KNN). To govern how well the model may be utilized to improve the accuracy of diagnosis of Heart Attack inside any patient, a very useful technique was applied. The suggested model's performance was quite pleasing, as it was possible to forecast evidence to confirm a heart illness in a particular person using KNN & Linear Regression (LR), with high accuracy when compared to Naïve Bayes.

Keywords: Machine learning, Classifiers, Heart disease prediction, KNN.

1 Introduction

The Global Health Analytics 2012 study illuminates the reality that one in five people globally had already brought up blood pressure – an illness that causes approximately 50 percent of all fatalities from heart attacks and strokes.

Cardiovascular Diseases (CVD), as well known as heart disease, encircles a range of criteria that impact the heart – not only heart conditions. In many nations, particularly India, cardiovascular disease was indeed the major cause of death [1,2]. In the U. S., 1 patient dies from cardiovascular

[1]Assistant professor, Department of CSE, VFSTR (Deemed to be university), Guntur, Andhra Pradesh
[2]Assistant professor, Department of CSE, Vignan's Lara institute of technology and science Guntur, Andhra Pradesh
[3]Assistant Professor, Department of Information Technology, P V P Siddhartha Institute of Technology, Vijayawada
***Corresponding Author:** syedshareefa@gmail.com

disease every 34 seconds. Cardiovascular disease includes coronary artery disease, arrhythmia, and heart illness. The term "heart disease" encompasses a wide range of conditions including narrowed or clogged blood vessels, and also the ways blood is pumped and circulated throughout the organ. A heart issue is a type of ailment that can be fatal. Every year, so many people are murdered by heart disease. Heart attacks are caused by the thinning of the heart muscle. The incapacity of the heart muscles blood fluid is commonly referred to as heart problems. A cardiovascular condition is also called CAD. A shortage of blood circulation to the artery may induce CAD [3-5].

Over the last decade, heart disease has surpassed cancer as the top cause of mortality worldwide (World Health Organization 2007). Heart problems, attacks, as well as other vascular illnesses are responsible for 41% of all deaths, according to the European Public Health Association [6]. Heart disease has a variety of symptoms, finding it challenging to identify it sooner and more accurately. Working with datasets of patients with heart disease is comparable to a real-world application. Doctors' expertise allows them to assign a numerical value to each characteristic. The trait with the greatest influence on disease prognosis is given equal importance. It also gives healthcare providers an additional source of information to help them make judgments.

Each year, thousands of citizens develop cardiovascular disease, but it is the first highest cause of mortal among patients throughout the United States and around the globe [7-10]. According to the World Health Organization (WHO), cardiac disorders cause 12 million people annually. Cardiovascular disease murders one individual every nearly 34 seconds around the world. Medical diagnosis is an important yet difficult process that must be completed quickly and precisely [11]. Data mining is a process of discovering heretofore unrecognized trends and patterns throughout datasets and using that knowledge to generate predictions. To collect hidden knowledge from huge data, data analysis refers to the statistical, machine learning, and database management system.

A substantial quantity of healthcare information is collected by the medical industry, which must be extracted for hidden data to make the appropriate decision. Research teams have used the data mining method to identify valuable knowledge from the massive number of clinician information available [12,13]. This has already been empowered by the declared open fatalities of heart patients every year and the accessibility of massive volumes of data on which to retrieve valuable knowledge.

Data mining is the technology of analyzing voluminous source information to recognize existing unseen trends, connections, and information that seem to be hard to identify using traditional statistical methods. As a result, data mining is used for extracting information from huge volumes of the data set. Data mining applications would be utilized to improve health policies and the avoidance of medical mistakes, illness early recognition, and avoidable hospital fatalities [14]. Depending on the medical patient information, a heart disease prediction technology can help healthcare practitioners in heart disease diagnosis. As a result, by deploying a data mining techniques system based on Mining techniques and performing a few types of mining on different cardiovascular disease variables, it'll be able to predict the patient's likelihood of being identified with cardiovascular disease greater accurately. This research introduces a new model that improves the accuracy of Decision Trees in detecting heart disease sufferers. It employs a decision tree method that is unique.

Diagnosis is a hard and vital activity that must be finished accurately and on time. The diagnosis is frequently developed based on the physician's skill and understanding. This leads to unfavourable outcomes and high medical expenses for therapies given to patients. Finally, an automatic medical diagnostics system would be highly advantageous. Our research goal is to give a complete verification of the many information mining approaches that might be utilized in such an automatic system.

2 Literature review

In medical centers, a percentage of work has been performed on illness diagnosis systems utilizing various machine learning techniques. Senthil Kumar Mohan et al.[15] suggested Effective Cardiovascular Disease Predictions Utilizing Mixed Machine Learning Approach, an approach whose goal is to uncover essential attributes using Machine Learning, hence increasing the prediction value in cardiac ailment detection. The expectations framework is made up of numerous few well-known arranging methods as well as a similar method. It is also informed regarding Different mining of information including expectations approaches, including as KNN, LR, Support Vector Machine (SVM), Neural Networks (NN), and Voting, have been popular of late to identify and forecast cardiovascular disease with only an acceptable accuracy of 88.7 percent using the predictive models for cardiovascular disease using a combination of various random forest methods with either a linear regression model. Finding of CVD Utilizing Machine Learning Techniques by Sonam Nikhar et al [16]. The focus of this research is to have a detailed description of the Naive Bayes & random forest classifications that we utilize in this study, particularly for predictive modeling. On the same dataset, more analysis was carried to see if a futuristic data gathering methodology could be used, as well as the results indicated that the Decision Tree algorithm excelled Bayesian different classifiers. The Multi-Layer Perceptron (MLP) neural network methodology was utilized to test and evaluate the data throughout this paper's suggested system.

This technique will have multiple levels, including an input level, an output level, and one or more hidden layers between both the two input levels. Each node in the input layer is connected to the external terminals. Each node throughout the input layer is connected to the data terminals by the neural network. This link is given a certain amount of weight. The biased input, which has a frequency of b, would've been added to the nodes to balance the convolutional layer. The relationship between all the two nodes might well be feed-forwarded or reinforced, depending upon the needs.

Abhay Kishore et al. [17] developed Heart Condition Prediction Using Deep Learning. This research establishes a heart condition prognosis method that includes Deep Learning techniques as well as an explicit Neural Internet Network to estimate the likelihood of a patient getting a heart-related ailment. Convolution Layer is a surface characterization algorithm that uses the Machine Learning method of Artificial Neural Network (ANN).

The article delves into the framework's primary modules and also the theory that underpins them. Big data and data mining are used in the proposed approach to deliver exact outcomes with the least faults.

This work serves as a foundation for the development of a new type of heart stroke detection technology. Machine Learning Methods for Predicting Heart Disease, Lakshmana Rao, et al., [5] wherein the significant components for cardiovascular disease are seen to be more (circulatory strain, current smoker, etc..).This article uses classification methods to analyze heart disease in adult patients. This article contains thorough information regarding Coronary Heart Diseases, including Facts, Common Forms, and Health Conditions. Waikato Environment for Knowledge Analysis, an excellent Data Mining Software for Bioinformatics Fields, was employed as the Data Mining technique. All 3 WEKA interfaces are being used here; the main data mining methods are Naive Bayes, ANN, and Random Forest, and the mentioned methods are employed to forecast cardiovascular attacks inside this computer. Decision Tree algorithms, such as CART, ID3 Algorithm, and Naïve Bayesian Approaches, are the most used methods for predictions shown in Figure 1.

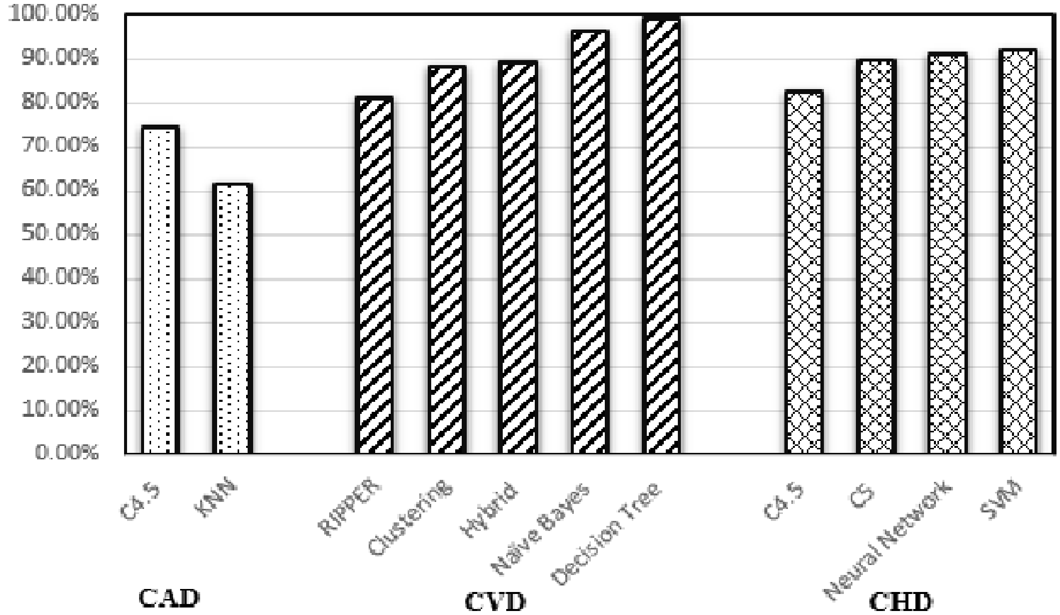

Figure 1 Comparison result of all classification for heart diseases

3 Methodology

For this cardiac prediction model, the following goals have been specified. The chosen system must've been scaled enough to execute against big databases with thousands of items, but it should not presume any previous information about the patient record keeping it is analyzing. This research examines different data mining strategies that can aid medical analysts and professionals in making accurate cardiovascular disease diagnoses. We initialize with a database which includes the medical histories of 294 people of various age group. This dataset gives us all crucial data, including the person's age, resting blood pressure, fasting blood glucose, and other clinical characters, that makes us in finding whether the patients have indeed identified with cardiovascular attacks. This dataset comprises 12 clinical variables for 294 individuals that allow us to determine whether or not the patient is a danger of developing heart problems, as well as categorize individuals who are at danger those who are not. This Cardiovascular Disease dataset was collected from Irvine's repository.

This leads to the identification of people who are at risk of a heart attack is extracted using this information. The training process is the two parts of these data. Every column is applied to a particular entity in this dataset, which comprises 300 rows & Fourteen columns. This research examines a variety of machine learning methods, including K neighbors (KNN), Logistic Regression, and Decision Tree Classification methods, all of which can aid clinicians & professional researchers in effectively diagnosing heart disease.

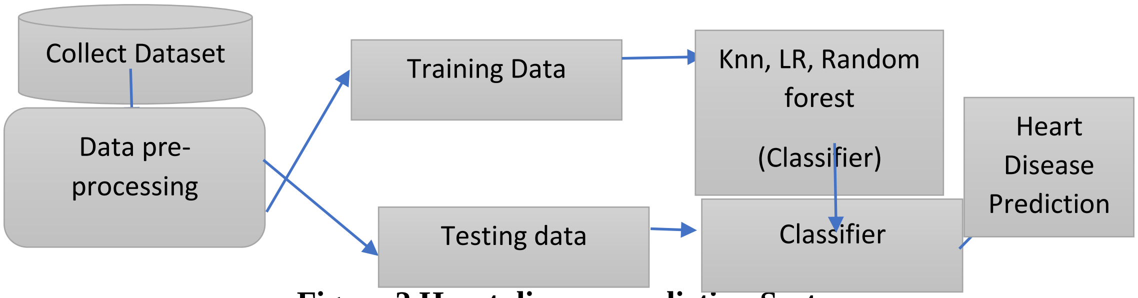

Figure 2 Heart disease prediction System

Checking publications, published studies, and data on heart diseases from recent years are all part of this process. The approach is a flow of procedures that transforms actual data towards recognizable trained data for people to understand. The proposed approach (Figure 2) consists of three phases, first is information gathering, next is the extraction of highest effects, and finally, the processing phase, in which we study the information. Depending on the procedures utilized information pre-processing tackles with null values, data analysis, and standardization. KNN, Logistic Regression, and Random Forest are the classifiers utilized in the proposed system to categorize the data once they have been pre-processed. Lastly, we put the suggested model to the test, evaluating it for performance and accuracy that used a variety of performance indicators. Using several classifications, and efficient Heart Disease Prediction Strategy has also been built in this framework. This system has been constructed employing classification techniques and uses 13 medical parameters including chest discomfort, fasted glucose, hypertension, cholesterol, age, and gender for disease prediction.

3.1 KNN

It is indeed a machine learning method and approach that could be used for linear and non - linear applications. KNN evaluates the categories of a certain amount of data surrounding a particular data set to forecast which category the data point belongs towards. It is among the most common algorithms because it is computationally efficient yet immensely powerful. Let's take a closer look at the KNN and see how it operates.

When a KNN algorithm is executed, it consists of three main basic stages:

- Set K to several clusters you want.
- The difference between the test example as well as the database instances is calculated then the computed distance is sorted.
- Obtaining the top K elements' descriptions and the forecast about the training sample is returned.

First, we'll decide on the number of neighbors, thus we'll go with k=3. The Euclidean distance in-between data sets will then be calculated. The Euclidean (E) is based on the distance among 2 points. It can be calculated using the following equation (1): Figure 3 shows the new data point can belong to category A because it has more nearest neighbors than B.

E (a,b) =$\sqrt{}$(u2-u1)^2 +(V2-V1)^2 ……………..(1)

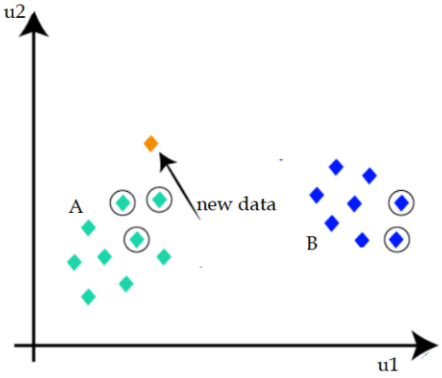

Figure 3 More neighbors from A so new data belongs to A

3.2 Logistic Regression

The classifiers in constructing logistic regression trees, which are made up of a tree-based architecture with such a logistic regression equation only at branches, are Logistic Model Trees. The approach can handle multi-class & sequential output variables, numerical & nominal characteristics, including lacking features. If there should be enough information to justify a more sophisticated tree structure, a mixture of learns which depend on simplistic linear regression when there is only limited and/or messy information then adds a much more complicated tree structure whether there is enough information to warrant such a framework. Cost-complexity thinning is used in the Logistic Model Tree (LMT). The speed of this method is much slower than that of the others. The evaluated attributes are connected with each inner node, just like in a decision tree.

LMT, like some other tree impelling devices, doesn't require any variable adjustment. LMT generates a single tree with double parts for numeric characteristics, multi-route portions for apparent characteristics, and linear logistic regression at the leaf, with the method ensuring that only relevant attributes are included in the final.

3.3 Random Tree forest

A random forest is a type of supervised classifier that comprises a large number of decision trees. Individual trees reflect the outcome of the categories. It is based on the random forest introduced by Bell Labs' Tin Kam Ho in 1995.

This strategy is used in conjunction with a random feature extraction to create decision trees featuring limited variability. The model is constructed using the specified methodology.

4. Result Analysis

Even though most researchers use diverse techniques including Decision tree, and Logistic regression to identify individuals suffering from Cardiovascular disease, the analysis shows that KNN, Random Forests, and Logistic regression outperform them. The techniques we utilized are much more precise, save a lot of money (i.e., they are cost-effective), and are quicker than that the techniques utilized by earlier studies. Furthermore, the prediction error attained by KNN and Logistic Regression is 81 percent, which would be stronger or nearly comparable to prior study levels of accuracy.

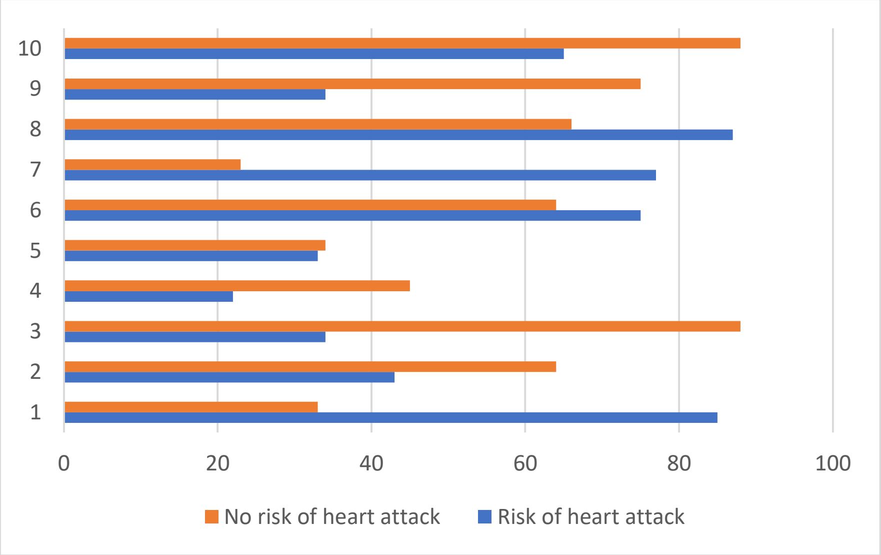

Figure 4: Age-based risk of heart problems

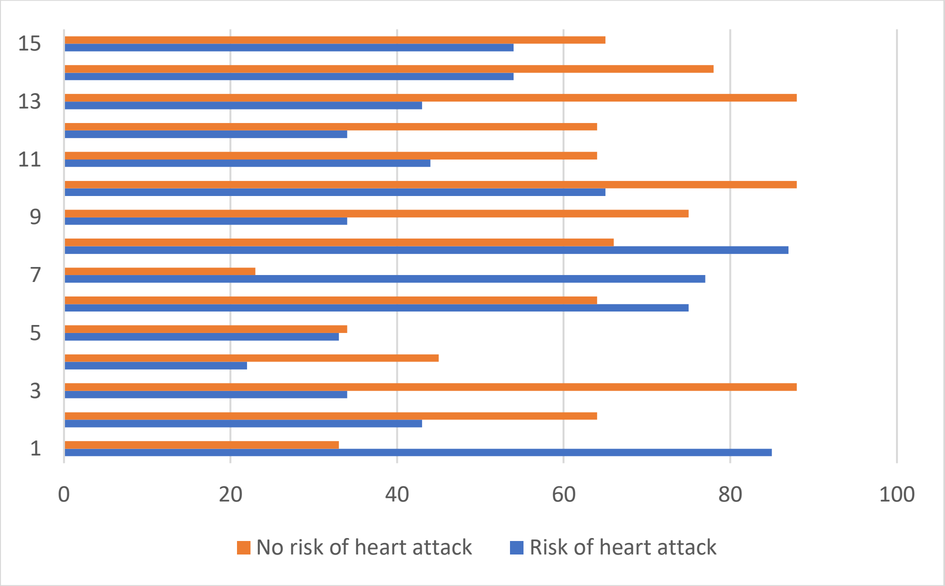

Figure 5: Blood pressure-based risk of a heart attack.

As a result of the enhanced medical information we used from information, we may conclude that overall reliability has increased. In the predictions of patients affected by heart illness, our experiment also shows that Logical Regression and KNN outperformed Random Forest Classifiers. This demonstrates that KNN and Logistic Regression are more effective in detecting heart problems. The plots in Figures 4 and 5 illustrate the volume of patients classified and predicted by classifiers based on age category and blood pressure.

5. Conclusion

A heart disease identification framework has been created using three Machine learning classifiers modeling approaches. This proposal forecasts individuals with heart disease by extracting the medical history which tends to lead to deadly heart problems from one dataset that contains patients' medical records such as heart problems, blood glucose, blood pressure, etc. This Cardiovascular Disease Detection Method aids a client solely on medical data from a previous heart disease diagnosis. Logistic regression, Random Forest Classifier, & K - nn are indeed the techniques utilized to create the provided model. As a result, this research assists us in predicting patients diagnosed with heart problems by cleansing the database and using logistic regression and K - nn to achieve an accuracy rate of 88 percent on our models, which is higher than that of the prior systems' efficiency of 85 percent. Furthermore, the accuracy of KNN, which is 88.4 percent, is the greatest of the three methods we used.

References

[1] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava ―Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques‖, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019 S.P. Bingulac, ―On the Compatibility of Adaptive Controllers,‖ Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994. (Conference proceedings)

[2] Sonam Nikhar, A.M. Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms" International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication,[Vol-2, Issue-6, June- 2016].I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Garikapati, P., Balamurugan, K., Latchoumi, T. P., & Malkapuram, R. (2020). A Cluster-Profile Comparative Study on Machining AlSi 7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon, 1-12.

[4] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD)," Prediction of Heart Disease Using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1- 5386-0965-1

[5] Abhay Kishore, Ajay Kumar, Karan Singh, Maninder Punia, Yogita Hambir," Heart Attack Prediction Using Deep Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04 | Apr-2018.

[6] Latchoumi, T. P., Vasanth, A. V., Bhavya, B., Viswanadapalli, A., & Jayanthiladevi, A. (2020, July). QoS parameters for Comparison and Performance Evaluation of Reactive protocols. In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE) (pp. 1-4). IEEE.

[7] Ezhilarasi, T. P., Kumar, N. S., Latchoumi, T. P., & Balayesu, N. (2021). A Secure Data Sharing Using IDSS CP-ABE in Cloud Storage. In Advances in Industrial Automation and Smart Manufacturing (pp. 1073-1085). Springer, Singapore.

[8] Sekaran, K., Rajakumar, R., Dinesh, K., Rajkumar, Y., Latchoumi, T. P., Kadry, S., & Lim, S. (2020). An energy-efficient cluster head selection in wireless sensor network using grey wolf optimization algorithm. TELKOMNIKA, 18(6), 2822-2833.

[9] A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar,” Machine Learning Techniques For Heart Disease Prediction”, International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019.

[10] Arunkarthikeyan, K., & Balamurugan, K. (2021). Experimental Studies on Deep Cryo Treated Plus Tempered Tungsten Carbide Inserts in Turning Operation. In Advances in Industrial Automation and Smart Manufacturing (pp. 313-323). Springer, Singapore.

[11] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025. [7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[12] Sneha, P., Balamurugan, K., & Kalusuraman, G. (2020, December). Effects of Fused Deposition Model parameters on PLA-Bz composite filament. In IOP Conference Series: Materials Science and Engineering (Vol. 988, No. 1, p. 012028). IOP Publishing.

[13] Yarlagaddaa, J., Malkapuram, R., & Balamurugan, K. (2021). Machining Studies on Various Ply Orientations of Glass Fiber Composite. In Advances in Industrial Automation and Smart Manufacturing (pp. 753-769). Springer, Singapore.

[14] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[15] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[16] Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[17] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.