Heart Attack Probability Analysis Using Machine Learning

Conference Paper · November 2021

DOI: 10.1109/DISCOVER52564.2021.9663631

6 authors, including:

Chinmai Shetty
NMAM Institute of Technology
8 PUBLICATIONS **42** CITATIONS

SEE PROFILE

2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) | 978-1-6654-1244-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/DISCOVER52564.2021.9663631

Heart Attack Probability Analysis Using Machine Learning

Annapurna Anant Shanbhag
NMAM Institute of Technology
Nitte-574110, Karnataka, India
annapurna.shanbhag99@gmail.com

Chinmai Shetty
NMAM Institute of Technology
Nitte-574110, Karnataka, India

Alaka Ananth
NMAM Institute of Technology
Nitte-574110, Karnataka, India

Anjali Shridhar Shetty
NMAM Institute of Technology
Nitte-574110, Karnataka, India

Kavanashree Nayak K
NMAM Institute of Technology
Nitte-574110, Karnataka, India

Rakshitha B R
NMAM Institute of Technology
Nitte-574110, Karnataka, India

***Abstract*- Heart Attack is one of the most common diseases observed in people of middle age as well as old age in the present day scenario. This may be due to unhealthy food habits and negligence of health in most people. Detecting the risk of heart attack and taking timely medication, can prevent serious illness. In this paper we explain about the different machine learning approaches and techniques used for predicting the probability of heart-attack risk. Different models are applied for heart-attack risk prediction. The probability of heart attack risk is displayed through a website. If a person is found having risk, suitable precautions are displayed under the guidance of the cardiologist. The proposed work analyses whether the person has a normal range of values for some highly contributing attributes which lead to heart attack like Cholesterol, Blood pressure, Blood sugar. The proposed work has better results compared to the previous work in terms of accuracy of prediction with highest value of accuracy as 85.7% for SVM model.**

Keywords- Heart Attack, passive, Nominal, Numerical Outliers, Noisy data, Null values, Attributes, EDA [Exploratory Data Analysis], Data mining, Machine-learning, Accuracy

I. Introduction

Heart-attack is one of the typical diseases in today's world. When the coronary arteries of the heart get blocked due to the deposition of cholesterol, fat and other substances the flow of blood to the heart gets blocked which causes heart-attack. Around 26 million people [1] in the world are affected by the risk of heart-attack. In the coming years, this ratio is expected to increase rapidly. Hence, we need to take some efficient precautions to reduce the risk.

The proposed work is to diagnose the risk of heart attack in a timely manner using machine learning technology. Dataset is collected from the Kaggle website. Data preprocessing is done in order to remove outliers and to correct missing and null value.

Five different machine learning algorithms are applied to train the dataset and their accuracies are obtained. The algorithm with highest accuracy is used to find the risk of heart attack. If a person is found to have risk, then suitable precautions are specified in a report which are suggested by the cardiologist. Dataset is analyzed to extract some useful information to make right decisions in the future.

This section gives an overview of risks involved due to heart-attack. Remaining parts of the paper are organized as follows. A brief overview of state-of-art literature is given in Section II. Proposed methodology along with relevant diagrams are given in section III. Section IV gives the experimental analysis along with the results and discussion. We conclude our work in section V.

II. Literature Survey

The work proposed in [1] attempts to improve the effectiveness of the classifiers by using various machine-learning models. The dataset was collected from a variety of medical databases. The researchers have used different techniques like decision tree, naïve bayes, random forest, logistic regression and SVM. In their study, SVM has given the highest accuracy with 92.30%.

The authors in [2] presented the adverse effects of Coronary HeartDiseases (CHD) and the necessity of having ML algorithms to predict the risk. They collected data from the Kaggle dataset that has 4000 tuples and 15 attributes. Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset. Support vector machine performed better, and they considered it in their project.

The work done in [3] predicts the heart attack risk for male patients only. It explains about Coronary Heart Diseases (CHD). That includes facts, common types, risk factors etc. The data mining tool that they have used is WEKA (Waikato Environment for Knowledge Discovery). Naive Bayes, ANN and decision tree are used as mining algorithms.

Authors have clearly explained about the heart attack and data mining tools used. It mentions the types of heart attack. The data mining techniques used in this work are Decision tree, artificial neural network and Naive Bayes[4].

The main aim of this paper [5] is to predict heart attack using Data mining algorithms. The main techniques used here are KNN and decision tree algorithms like CHAID, CART, C4.5, J48, ID3 etc., and naive bayes algorithms.

This paper deals with a few machine learning techniques to predict heart disease. Data preprocessing is carried out to treat outliers and missing values. They have used Naive Bayes, K means and decision tree in this paper. Among these models it is found that Naive Bayes provides accurate results. [6].

In this paper, a Naive Bayes classifier is used which is based on Bayes theorem. Hence it makes assumptions independently. The dataset had 500 patients collected from a diabetic research institute of Chennai. The Weka tool is used here [7]. The accuracy produced by Naive Bayes is 86.42%.

Miranda has used the Naive Bayes classifier to predict heart disease. This approach produced 85% accuracy when applied on a dataset [8].

In this paper [9] authors talk about the analysis that helps to predict the health condition of the user using normal scan and machine learning techniques. In the present day where people want their work to be done as fast as they can by using a laptop, mobile, or any other device, this analysis provides a solution to build a mobile app which can produce medical reports rapidly.

The k-Nearest Neighbors (KNN) is one of the effective methods for classification [10]. This algorithm works on a given K value. We can choose the K value in many ways but a simple one is to run the algorithm on different K values many times and choose the one which gives higher accuracy.

In this work, SVM algorithms are classified into three types, namely: Decomposition based algorithms, Variant based algorithms and others. This paper details the survey on various training algorithms which are mainly used for support vector classifiers [11].

In this work, authors propose ID3 algorithm [12] to predict heart disease. This method not only lets the patients know about their heart disease but also helps to reduce the death rate. This also keeps the count of people affected with heart disease.

The work presented in a paper [13] explained about MAFIA (Maximal Frequent Itemset Algorithm) and K mean clustering. For prediction of diseases classification is an important step. They used classification based on MAFIA and K Mean clustering. The results were accurate.

In this proposed work the K star algorithm [14] was used to diagnose the level of coronary heart disease. They have utilized Learning vector Quantization neural system calculation. It predicts the nearness of illness in the patients by considering various parameters.

Authors exhibit a framework for heart infection in the paper [15]. They have used Learning vector quantization neural network system calculations.

In [16] the dataset is chosen from UCI repository. The proposed work predicts whether patients have heart disease or not. Here 14 attributes are considered. To train three algorithms Logistic regression, KNN and Random Forest Classifier are used. KNN was found to be the best suited for the scenario with an accuracy of 88.52%.

Data is obtained from the UCI repository which consists of 303 patient records, where 6 records were found to have some missing values were removed as a part of preprocessing. ML models like decision tree, random forest, language model, support vector machine is used, and performances are compared [17].

Authors have proposed their work to predict heart disease at an early stage using data mining [18]. Data source is the Cleveland heart disease data set. Various machine learning models are used like Naïve bayes, decision tree and SVM. The accuracy obtained is 84.1%.

Various classification techniques were used to build risk prediction models for heart disease. Data source is the Cleveland heart disease data set. The techniques used are decision technique, association rule, KNN, ANN, naïve bayes and hybrid approach. The accuracy obtained is 83.66% [19].

Comparative study and analysis of data mining classification methods are used in cardiovascular disease prediction. The data source is cardiovascular disease dataset. SVM, artificial neural network (ANN), decision tree and ripper classifier are used, and accuracy obtained is 84.7% .

III. METHODOLOGY

In this section the methodology of the entire project is clearly explained. Fig 1 shows the sequential chart of our proposed work. Steps are from collecting the data to predicting the heart attack risk and suggesting the precautions. The details of all the steps are thoroughly explained in the upcoming sections. The dataset used here is known as Heart Disease Dataset taken from Kaggle website. It has 14 attributes with 303 patients' records. The attributes mentioned here are useful to predict the risk of heart attack.

A. Dataset Description:

In this subsection, we discuss the attributes present in our dataset. We have listed a total of 14 attributes along with their description and distinct values. Table I gives the details of attribute description.

B. Data pre-processing:

Data preprocessing is a crucial step used to clean the data and experiment with machine learning or data mining. As the data collected from websites cannot be directly processed by the machine learning models, data pre-processing has to be done. This process is termed as Exploratory Data Analysis (EDA). In this process noise (meaningless data), outliers, inconsistent-data, null values, missing values were processed. As the models operate only on numerical data, nominal data must be

converted. At the end of this process standard dataset suitable for further processing is obtained.

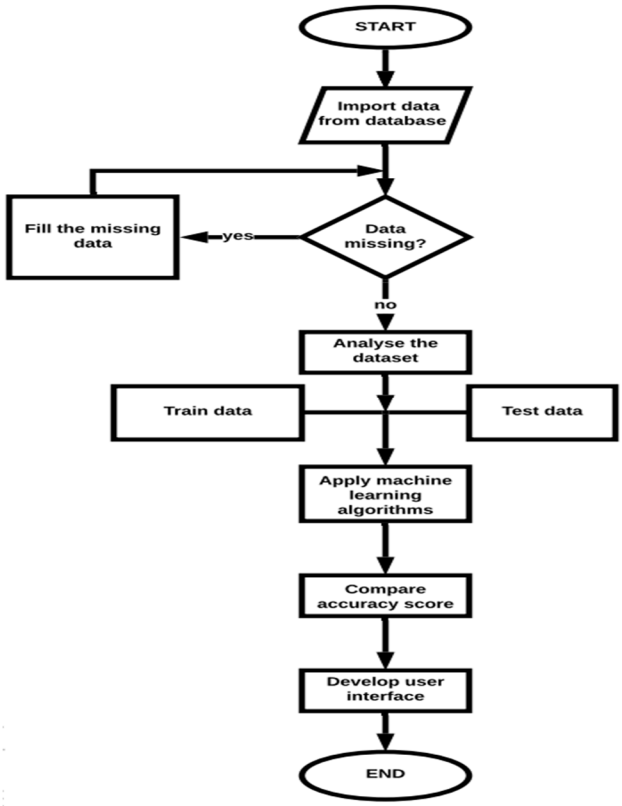

Fig. 1.Sequential chart of proposed work

TABLE I. DESCRIPTION OF ATTRIBUTES

S.No	Attribute Description	Distinct Values
1	**Age**– This attribute defines the age of the person.	
2	**Sex**- This attribute describes the gender of a person. [0 means Female and 1 means Male]	0,1
3	**CP** – This attribute defines the level of chest pain (CP) a patient is suffering from. There are 4 types of values. 0: Asymptomatic, 1: Atypical angina, 2: Pain without relation to angina, 3: Typical angina	0,1,2,3
4	**TrestBP**– This attribute gives the blood pressure (BP) of the patient.	
5	**Chol**– This attribute shows the serum cholesterol level.	
6	**FBS** – This parameter describes the fasting blood sugar level. If the patient has the value greater than 120mg/dl then 1 if not then 0.	0,1
7	**RestECG –** This attribute shows the result of ECG from 0 to 2. Each value shows the severity. [0-Probable left ventricular, 1-Normal 2-Abnormalities in the T wave or ST segment.]	0,1,2
8	**Thalach**– This attribute indicates the maximum value of heartbeat recorded at the time of stress.	
9	**Exang -** This attribute states whether exercise induced angina or not. [1-yes, 0-No]	0,1
10	**OldPeak –** This parameter defines the patient's depression status.	Real number values between 0 and 6.2
11	**Slope –** This parameter describes the patient's condition during peak exercise. [0-Down sloping, 1-Flat, 2-Ascending]	0,1,2
12	**CA**: This parameter shows the status of fluoroscopy.	0,1,2,3
13	**Thal -** This parameter describes the blood flow observed. There are four kinds of values. [0-NULL, 1-Fixed defect, 2:Normal blood flow, 3-Reversible defect]	0,1,2,3
14	**Target**- This attribute depends on all the 13 attributes mentioned above. There are two different types of values where 0 indicates there is a risk of heart attack and 1 indicates there is no risk.	0,1

C. Data Analysis

Data analysis has been done after data preprocessing to extract useful information. This information helps to make appropriate decisions in the future. Below graphs are plotted

for the attributes using some python commands based on the data obtained from EDA.

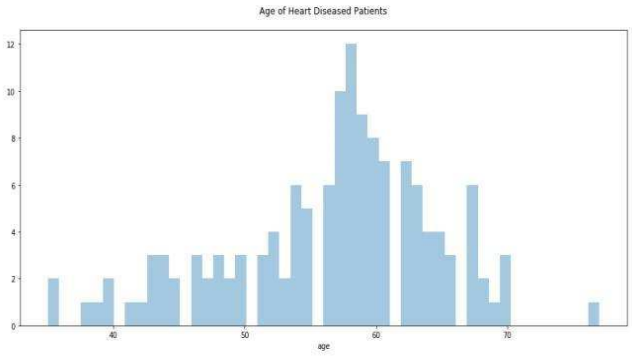

Fig. 2.Displot of attribute age

Fig. 2 depicts the distplot of age attribute. It clearly shows that heart disease risk is very common with the people whose age is 60 and above. The heart disease risk is common with the people who belong to the age group of 41 to 60 and rare with the people whose age is below 40.

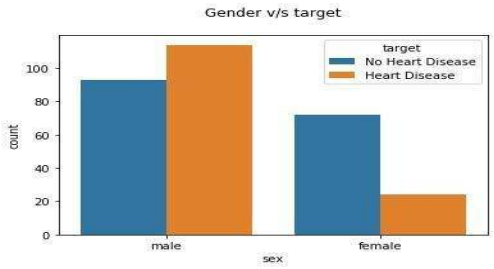

Fig.3. Countplot of sex

Fig 3 depicts the countplot of sex attribute. This graph shows that men are more likely to have a heart disease than women

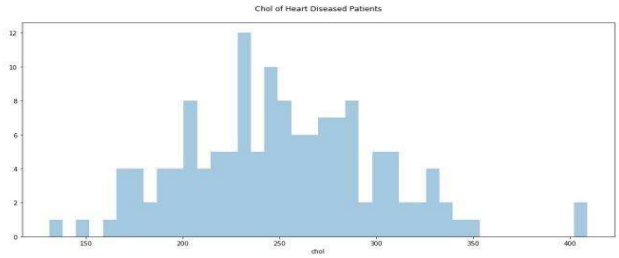

Fig.4. Distplot of chol

Fig.4. shows the distplot of cholesterol. In grown-ups, the total cholesterol levels are considered desirable less than 200 milligram per deciliter (mg / dL). Higher the cholesterol level, higher is the risk of heart diseases.

Fig 5 depicts the count plot of the target variable which has the values 0 (risk of heart attack) and 1 (no risk of heart attack). Around 140 records are collected from people who do not have the risk of heart attack and greater than 160 records have the chance of getting a heart attack.

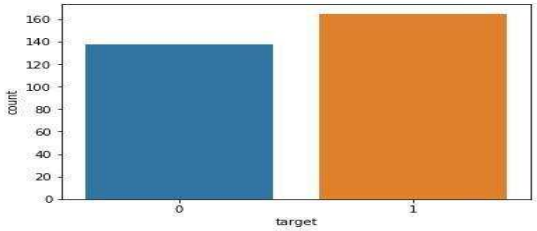

Fig. 5.Countplot of Target

D. Model selection:

This project uses five Machine Learning Models which are KNN, Decision Tree, Random Forest, Logistic Regression and SVM.

1) K Nearest Neighbor Model:

The k-Nearest Neighbors (KNN) is one of the effective methods for classification. This algorithm works based on a given value of K. We can choose the K value in many ways. But a simple one is to run the algorithm on different K values many times and choose the one which gives higher accuracy.

2) Decision tree:

The Decision Tree is a machine learning technique which has tree-like structure. Internal node represents the dataset attributes, and the outer branches represent the results.

3) Random Forest:

Random Forest algorithms are used in classification and regression. It forms a tree for the data taken and based on that makes a prediction. Random Forest algorithm usually works on large datasets.

4) Logistic regression:

Logistic Regression is a classification algorithm. It is used for binary classification problems. This algorithm makes use of the logistic function for squeezing the output of a linear equation between 0 and 1 instead of fitting a straight line or hyper plane,

5) Support Vector Machine:

SVM is one of the best classification techniques which achieve excellent performance in numerous applications. There are three algorithms. They are Decomposition based, Variant based and others. We should choose them according to their flexibility for specific applications.

E. Model Implementation and Performance Measurement:

The dataset obtained after Exploratory Data Analysis is followed by splitting data into training and testing sets to build the machine learning model. It is trained using training sets and tested using testing sets. And then the above-mentioned models are applied to find out which gives the highest accuracy.

IV. Results and Discussion

A. Confusion matrix

The model performance in the form of confusion matrix is shown below in table II.

TABLE II. ANALYSIS OF MODEL PERFORMANCE THROUGH CONFUSION MATRIX

Decision Tree	True(1)	True(0)
Prediction(1)	30	14
Prediction(0)	13	34
Random Forest	True(1)	True(0)
Prediction(1)	37	7
Prediction(0)	12	35
Logistic Regression	True(1)	True(0)
Prediction(1)	34	10
Prediction(0)	4	43
SVM	True(1)	True(0)
Prediction(1)	35	9
Prediction(0)	4	43

A confusion matrix is represented as a table which describes the performance of a classifier that is executed on the given test data, In the confusion matrix, the "True" values are known data values.

There were 303 patient records in the given dataset. For example, KNN is showing that 25 patients in the dataset had heart attack risk and predicted correctly. But initially 4 records belonged to heart attack risk but under the category (0) predicted wrongly; non-heart attack patient. In the same way, 39 patients without heart attack risk were predicted correctly whereas 8 patient's predictions were wrong.

Fig. 6 shows the comparison of accuracy for different models. In the proposed work, Support Vector Machine has given the highest accuracy (85.7%) whereas Decision Tree has the least accuracy (70.33%) of all the models.

Comparisons of all models are in Fig 6.

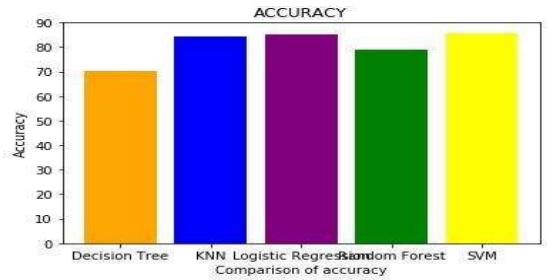

Fig.6.Accuracy comparisons of models.

Table III gives the accuracy achieved by each of these models.

TABLE III. ACCURACY OF MODELS

Model	**Accuracy**
KNN	84.2%
Decision Tree	70.33%
Random Forest	79%
Logistic Regression	85%
SVM	85.7%

B. Developing User Interface:

After finding the accuracy of all models and selecting the best model (SVM), we are showcasing its application in the real word by creating a website. The website has 3 major parts. They are: Front end, Back end, Database.

Front end**:** For front end creation we have used React which is one of the most popular and trending JavaScript libraries. Along with react core features like redux, hooks, we have also imported other libraries for our task competition. Our front end is running on localhost:3000.

Back end: We have used python Flask as backend to implement the SVM model. It is running on localhost:5000. Front end and backend are connected through API calls.

Database: We are storing user data in Firebase real-time database for future. It's a cloud hosted NO SQL database. It stores data in key- value pairs.

C. Registration and usage of the website:

Firstly, user must register or log-in to use the website. He has to fill his personal information such as name, gender, date of birth, phone number, email and password to get registered. This information will be stored under 'userDatabase' in Firebase. Later he should fill all the values for heart-attack related attributes which are mentioned above (section 3.1). This information will be stored under 'attackInfoDatabase' in the Firebase. Later these attribute values will be sent to Flask where data will be processed by SVM. Based on that either 0 (Heart attack Risk is found) or 1 (Heart attack Risk is not found) will be sent back to react. This exchange of data between React and Flask takes place through API calls. Based on the value sent by the Flask, the result will be displayed to the user.

A detailed report will be generated in pdf format.

Generated report contains

- User's personal information.
- User's medical information related to heart attack.
- Graphical analysis of Cholesterol content (chol), resting blood pressure level (trestbps) and fasting blood sugar (fbs)
- Pecautions are also listed for those above mentioned highly contributing attributes. These precautions displayed under cardiologist.

Figures 7 and 8 depict the user interface of the model.

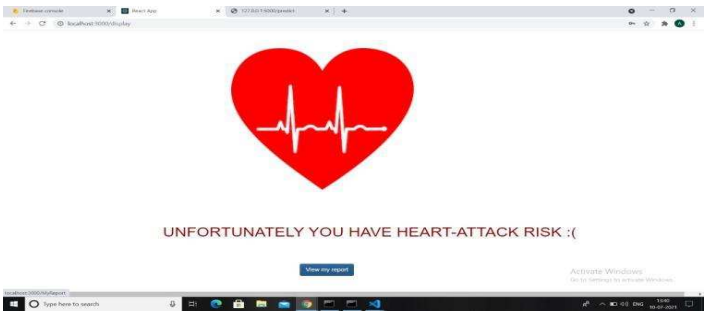

Fig. 7. Output page (Frontend)

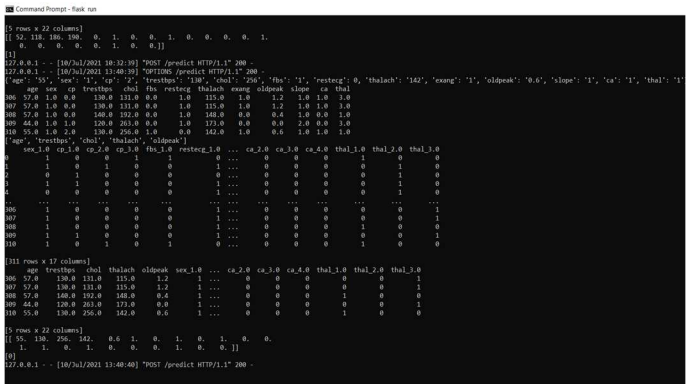

Fig. 8.Flask command prompt (Backend)

V. CONCLUSION

In this project we have discussed the various models used for predicting the heart-attack risk. We found their accuracy and SVM gave the highest accuracy 85.7 %. So SVM was used to find the heart attack risk. This project will help the patients to know whether they have chances of getting heart attack or not based on their medical history without consulting the doctor. If the risk has been found, it will be displayed to the users along with the precautions to be taken. Also, the dataset is analyzed to draw some inferences so that they can be used in the future to make right decisions.

The dataset used in this project has only 303 patient records and hence the result displayed to users might not be accurate. This is the main limitation of the carried-out work and hence the datasets need to be amplified for accurate results. Heart-attack history of the patients (how many times they had heart attack in the past) is one of the most important attributes for heart attack risk prediction which is missing in the dataset. For 'sex' attribute only two genders are considered. Transgender category should be considered along with male and female category in the dataset. Along with machine learning models, deep learning models can also be used for better accuracy and accurate results.

REFERENCES

[1] F. S. Alotaibi1 The International Journal of Advanced Computer Science andApplications,Vol. 10, No. 6, 2019

[2] A. Mordecoi, "Heart attack risk prediction using machine learning", April 2020

[3] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). IEEE, Feb. 2015. doi: 10.1109/ablaze.2015.7154917.

[4] A . Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," 2017 International Conference on Intelligent Computing and Control (I2C2). IEEE, Jun. 2017. doi: 10.1109/i2c2.2017.8321771.

[5] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," Computer Science & Information Technology (CS & IT). Academy & Industry Research Collaboration Center (AIRCC), Aug. 08, 2014.

[6] N. Rajesh, M. T, S. Hafeez, and H. Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," International Journal of Engineering & Technology, vol. 7, no. 2.32. Science Publishing Corporation, p. 363, May 31, 2018. doi: 10.14419/ijet.v7i2.32.15714.

[7] D. Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal,Vol. 5: Iss. 2, Article 1.,2014

[8] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detecof Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier," Healthcare Informatics Research, vol. 22, no. 3. The Korean Society of Medical Informatics, p. 196, 2016.

[9] C. Thirumalai, A. Duba, and R. Reddy, "Decision making system using machine learning and Pearson for heart attack," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA). IEEE, Apr. 2017. doi: 10.1109/iceca.2017.8212797.

[10] G. Wang, "A Survey on Training Algorithms for Support Vector Machine Classifiers," 2008 Fourth International Conference on Networked Computing and Advanced Information Management. IEEE, Sep. 2008. doi: 10.1109/ncm.2008.103

[11] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, pp. 986–996, 2003. doi: 10.1007/978-3-540-39964-3_62.

[12] P. Athilingam, B. Jenkins, M. Johansson, and M. Labrador, "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial," JMIR Cardio, vol. 1, no. 2. JMIR Publications Inc., p. e3, Aug. 11, 2017. doi: 10.2196/cardio.7848.

[13] D. Abd, J. K. Alwan, M. Ibrahim, and M. B. Naeem, "The utilisation of machine learning approaches for medical data classification and personal care system mangementfor sickle cell disease," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT). IEEE, Mar. 2017.

[14] M. Shouman, T. Turner, and R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," International Journal of Information and Education Technology. EJournal Publishing, pp. 220–223, 2012. doi: 10.7763/ijiet.2012.v2.11

[15] J. Amudhavel, S. Padmapriya, R. Nandhini, G. Kavipriya, P. Dhavachelvan, and V. S. K. Venkatachalapathy, "Recursive Ant Colony Optimization Routing in Wireless Mesh Network," Advances in Intelligent Systems and Computing. Springer India, pp. 341–351, Sep. 11, 2015. doi: 10.1007/978-81-322-2526-3_36.

[16] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1. IOP Publishing, p. 012072, Jan. 01, 2021. doi: 10.1088/1757-899x/1022/1/012072.

[17] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, no. 6. Springer Science and Business Media LLC, Oct. 16, 2020.

[18] N. K. S. Banu and S. Swamy, "Prediction of heart disease at early stage using data mining and big data analytics: A survey," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). IEEE, Dec. 2016.

[19] G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease," Indian Journal of Science and Technology, vol. 8, no. 12. Indian Society for Education and Environment, Jun. 24, 2015. doi: 10.17485/ijst/2015/v8i12/58385