Heart Disease Prediction Using ML Algorithms

Apoorva Kyramkonda
Computer Science, University of Central Missouri, USA

Sanugommula Akshitha Reddy
Computer Science, University of Central Missouri, USA

Vakrkuti Vijay Bhasker Reddy
Computer Science, University of Central Missouri, USA

Pasula Sai Ragasri
Computer Science, University of Central Missouri, USA

Abstract:

These are of the most common reasons for death within the contemporary age is heart disease. The medical analysis process faces a significant problem when predicting heart disease. There has been demonstrated that machine learning (ML) can be useful for helping for generating decisions and forecasting again from a huge amount of information generated either by the medical sector. Additionally, we observed their employment of ML approaches in the latest advances across several IoT domains (IoT). If foreseen far ahead, such data could provide valuable details towards clinicians, allowing them to customize both prognoses but also courses of care for everyone. Quite a few research have looked at using ML can forecast cardiovascular problems. Throughout this study, we suggest one unique approach to improve the overall precision of CVD prognosis through identifying major characteristics using machine learning algorithms. This health system gathers a lot of information, yet sadly, not every bit of it needs to be processed to find undiscovered emerging trends and make wise decisions. We would examine heart disease forecasting models with a larger range of variables. Forecasting the increasing development of major cardiac illnesses is still a significant task that necessitates mostly in the healthcare profession. Machine learning could considerably simplify those millions of health services complications which have down nowadays to remaining unsolvable. This management information system for the treatment of pulmonary illness is indeed the subject of the project proposal. The research provided utilization of input data from the OpenML library that contains 1 million occurrences in cardiac illness and 14 characteristics. Machine learning methods such as random forest, decision trees, gradient boosted trees, linear support vector classifier, logistic regression and recurrent neural network are utilized to undertake executables but also multiclassification mostly on input data within a week of implementing preprocessing but also highlight methodologies. Inside this research, we utilize machine learning techniques to forecast probable cardiovascular disease within patients.

Keywords: Machine Learning, Heart disease, cardiovascular, IOT, XGBoost, Random Forest.

Github Link: https://github.com/ApoorvaKyramkonda/PROJECT

Introduction:

In accordance with WHO statistics, heart condition causes 17.9 deaths worldwide yearly, making this the world's top source of death. Among biggest behavioral cardiovascular diseases (cvds including strokes include poor diet, inactivity, cigarettes, but also alcohol intake. Another cardiac arrest happens whenever arterial artery walls impairs your brain's ability to pump blood. The strokes is brought caused by a thrombosis in such an arterial that restricts cardiovascular system. The effects are related to those experienced by other disorders and could be mistaken the signs of aging, rendering a prognosis challenging especially medical professionals. To enhance patient life expectancies, accurate cardiovascular illness prognosis with early detection were crucial. This growing gathering of medical information has given professionals a fantastic chance to advance hospital diagnostics.

About lot of time and money, physicians' assessments of a clinical encounter, symptoms, including records from clinical examination are applied in the existing approaches towards forecasting but also assessing cardiovascular problems. Various tests concerning individuals were already generally available through database inside the medical industry, which are growing every day. The information from the UCI ML collection was applied to this work on creating a forecasting model predicting high cholesterol. Utilizing the characteristics that seem to be included in the dataset, this same computer is trained to recognize connections. A successful machine learning strategy for predictions involves categorization. Classification would be an efficient supervised ML technique towards diagnosing diseases whenever professionally trained without reliable evidence. This same main objective of such an effort is always to build a medical heart attacks prediction system using modern Ml methods.

Nevertheless, while healthcare industry was somewhat accessible during important times, a few situations during which there is a paucity in supplies in such an emergency. Mostly in diagnostic of disorders such coronary heart disease, each tracking feature. Considering cardiac facilities including OPDs produce huge volumes of information linked to that same prediction of diseases, there is a tremendous opportunity using big data analytics that improve cardiac medication adherence results. Nevertheless, something is difficult to arrive at detailed, accurate, and consistent conclusions utilizing such input ambient noise, missing altogether, but also measurement errors. Owing with significant technological developments, preservation, acquiring, especially information restoration, machine learning is already a major factor in cardiologist. In addition to making choices employing different algorithms, experts had normalized information utilizing range of data mining algorithms.

The above work emphasis on research and development of such a decision support tool that forecast cardiovascular disease utilizing Fourteen features of medical studies. The literature review reviews all relevant studies conducted to date. Our suggested study highlights a specific appropriate method for precisely diagnosing that condition but also analyzes apparent gaps throughout the prior studies. Using data analysis, Methodologies but instead Outcomes offers classification methods. Overall assessment, resolution, but also

effectiveness of machine learning methods which may well successfully detect cardiac problems using medical evidence are presented. Throughout summary, overall effectiveness, evaluation, including contrasts from several kinds of techniques just on systems are presented.

(a) Motivation and Contributions:

Another primary reason for death throughout history was indeed cardiovascular disease. Cvds are indeed the leading cause of death worldwide, according to World Health Organization. killing 17.9 million population each year. Cardiovascular disease, idiopathic intracranial, aortic stenosis, and several illnesses comprise the cardiovascular and renal diseases collectively referred to here as CVDs. One among Incident cases over about of 70, while strokes and cardiac events account for 4 of each 5 fatalities. This planned production's major determinants would be as follows:

(i) Our first tackle the issue concerning information, before proceeding to further develop but also regulate. This is a significant leading contributor of the increasingly considering. This information is then utilized to train and evaluate these classifiers as identify those that always have higher precision.

(ii)This same optimal combination but rather characteristics are therefore chosen utilizing autocorrelation.

(iii)This final phase involved fine-tuning model settings using resampled database and machine learning techniques that accurate results feasible.

Cardiovascular disease can be predicted using an illness forecasting models. Additionally, this same objective of the study would be to figure out the ideal categorization system capable of detecting cardiovascular problems inside an individual. That while these machine learning techniques have been widely utilized, predicting cardiovascular dysfunction seems to be an important step requiring that accuracy and reliability. Consequently, a variety of grades but also assessment strategy categories have been employed to analyze these optimization techniques. It'll also assist scientists and healthcare professionals improve their understanding of the situation at hand but also point us inside the direction of both the right strategy towards forecasting cardiac problems.

Main Contributions & Objectives:

(i) The important scientific project's major finding optimal classification method that will offer the highest level of precision whenever classifying ordinary and unusual people.

(ii)To anticipate an individual's vulnerability to cardiovascular problems but also strokes as accurately as feasible which used a healthcare collected data.

(iii)Objective is to evaluate heart disease forecasting models utilizing larger range of input parameters. This algorithm analyzes 13 variables, including clinical words including Gender, Age, Blood Pressure, and Cholesterol, which forecast overall risk that a person would develop heart disease.

(iv)Many of the most difficult issues confronting the medical industry today was its ability to predict cardiac illnesses.

(v)Lastly, using a completely cost-effective strategy, we group patients according to whether they would be at likelihood of developing a medical problem or otherwise.

Literature Review:

There were several research but also techniques concerning how to categorize cardiovascular problems employing data mining and artificial intelligence.

In-depth study of the study upon that utilization of machine learning mostly in field treating cardiovascular problems was offered by Al-Janabi [13]. As stated by the writer, a database containing sufficient examples but also correct information is required to build an efficient algorithm regarding disease prediction. Preprocessing your information correctly is important because it may affect whether effectively the machine learning method utilizes actual information.

Marimuthu et al., suggested a cardiovascular disease estimation accuracy This publication includes discusses that application's effectiveness as well as a summary of earlier research.

Yadav et al., offered an information preparation infrastructure to be employed prior developing and testing different techniques. The incorporation of Classifiers by the writer to improve each ML individual's presentations was stressed. The study believes that variable adjustment was an effective method for achieving high accuracy.

Sharma et al., With the Collected the information of cardiovascular, it was suggested to employ a deep learning methodology to identify cardiovascular disease. Researchers stated because one of the important areas wherein deep neural networks could be utilized to improve the accuracy of classification is during the diagnosis of heart disease. Researchers were able to demonstrate how Talos hyperparameter optimization outperforms conventional alternatives for modeling.

Latha and Jeeva et al., concentrates upon using it with a database of health data. Overall results of the study demonstrate how clustering algorithms, including bagging and boosting, were beneficial for enhancing overall reliability of classification algorithm also excel for estimating the risk of chronic diseases. Incorporating classification model significantly enhanced the procedure' efficiency, or the outcomes showed a significant improvement in model accuracy. Poor classifiers benefited from ensemble classification, which increased efficiency by approximately to 7%.

Ravindhar et al. employed 1 computational model but also four machine learning methods that correlate quality measures to the detection of heart problems. These researchers measured their programs' reliability, specificity, memory, overall F1 parameters through order to determine their ability to forecast cardiovascular events. This deep neural networking system identified cardiac illness with 98% efficiency.

Guru, et al., have suggested expanding a decision - making framework that discover five important cardiac disorders using a computing design based on something like a multilayer perceptron having three levels. A

training algorithm method enhanced with said impetus component, this same innovative teaching speed, and even the remembering mechanism used to develop any suggested recommender system.

Noh, et al., another association classifiers built using the effective FPgrowth has been proposed as a classification technique. Researchers provided a criterion to quantify this cohesiveness but also, throughout turns, recommended a classification technique, as this is the construction of such an association classifier, since the quantity of designs could be quite varied but also large. Researchers provided other criteria that quantify cohesiveness since the quantity underlying layouts could be quite wide and varied.

McPherson et al., This integrated execution mechanism also used Neural Network approaches, that were only barely effectively capable of identifying yet if the testing subject seemed to have the specified ailment or otherwise.

Ganna A, Magnusson P K et al., a calm this research has indeed been significantly influenced by studies on applying machine learning methods that diagnose coronary heart problems. An overview of the research is provided in just this essay. And to use a variety of methods, an effective forecast of metabolic syndrome has indeed been achieved. Logistic Regression, KNN, Random Forest Encoder, etc. are a few of these. Findings demonstrate that every technique is capable of registering fulfilling specific goal.

Proposed Framework:

Taking the steps outlined in Fig. 1, particularly details the scientific approach used to develop the classification algorithm needed for clinical predicting heart disease, will enable you to forecast cardiovascular disease. This algorithm is a crucial step in whatsoever machine learning method used to forecast cardiovascular problems. A classifier must first be taught only with data in to create a classification algorithm, before being supplied with both a fresh, unidentified data to generate its forecast.

All four classification methods' efficiency was assessed utilizing 10-fold cross - validation but also percentile splits being part of the study's approach used in this study. Within cross validation, all cardiovascular disease training and testing data is typically divided using just several folds, like 10, but every folding being utilized repeatedly in both training and evaluation via replacing other folds throughout the database.

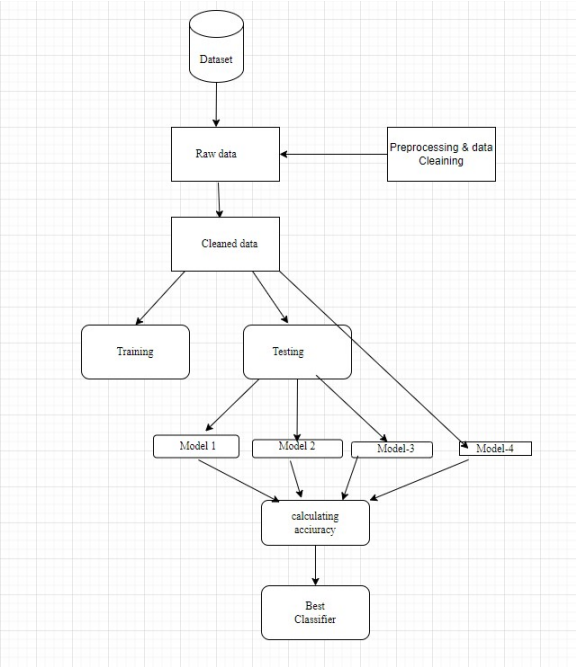

Fig: Flow chart of methodology

Decision Tree:

A decision tree is indeed a graph which employs a branch technique can show each potential result of something like a choice. Drawing tree structure via manually, using a graphical tool, as well as using sophisticated software have all been options. Whenever a company is confronted with a decision, decision trees can help concentrate the conversation.

To streamline decision-making, they can dynamically provide a large amount, temporal, and perhaps other quantities towards potential results.

Using a decision tree like a regression analysis as well as classification technique is a machine learning strategy. Each category has a tree-like organization. According to the picture following, this builds a tree of strategic options where the channel's nodes serve to represent the information, whose links represent together for classifications, but each leaf node, should one be chosen, serves to represent the intended classifier. A decision tree's category forecasts will look something like follows:

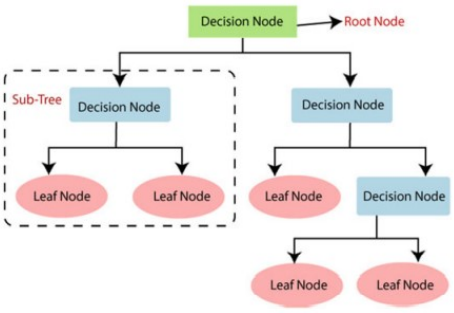

Fig: Architecture of decision tree

Following the eligibility requirements, this process merely starts somewhere at cluster head of both the choices trees; this then chooses other nodes yet continues to maintain ultimate leadership networks, whose classifiers resides only at the leaf node and appears to have been successfully classifying incoming categories.

Random Forest:

Random forest is a component of ML models. This emphasizes the concept of clustering methods, a technique that combines various categories to answer a challenging topic all while improving classification performance. As seen here, supervised learning is divided into n categories; nevertheless, each grouping is trained with a tree structure which creates n models. Polling was employed during the prediction, with the group receiving overall majority of votes being considered their final forecasted classification.

A supervised teaching method called Random Forest utilizes the ensembles reinforcement learning towards prediction. The supervised learning research introduces forecasts across various machine learning algorithms continue providing forecasts that are significantly reliable than those generated by a unified framework.

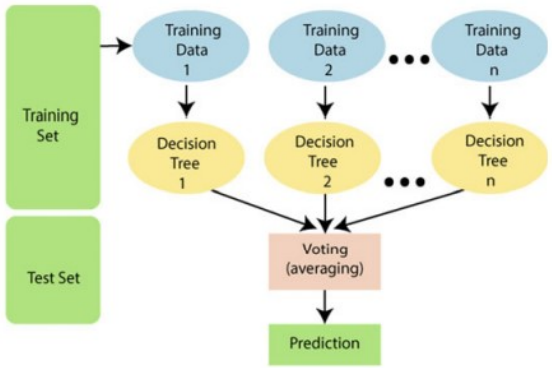

Fig: Architecture of Random Forest

SVM:

Seeing as it provides most best fit line that could partition n-d space into sub bands and the effective decision border, it's indeed helpful to pinpoint how well a new record belongs to just that categories. That vector has the designation of such a hyperplane. That linear vector support machine chooses the much more extreme examples that create the best lines. Likewise, these extreme places were known as support vectors. Inside the image below, each decision threshold that was utilized to classify this data can be seen.

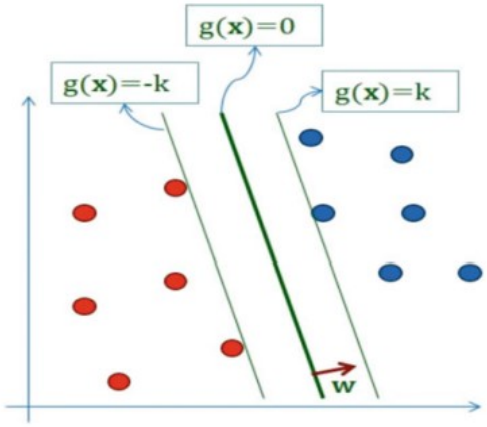

Fig: Hyperplane in SVM

XGBoost:

Two factors led to using an XGBoost classification tree:

1) the framework is constructed by dividing on either a particular trait,

2) it might be more resistant to certain other types of decision tree models.

However, there were just 14 characteristics as well as everything was considered as an independent variable to figure out which characteristics contributing towards the forecast, no feature selection technique has been performed. Additionally, I looked for any cointegration among some of the 14 parameters that confirm these conclusions. According to Pearson's correlation, there was no significant link itself between parameters.

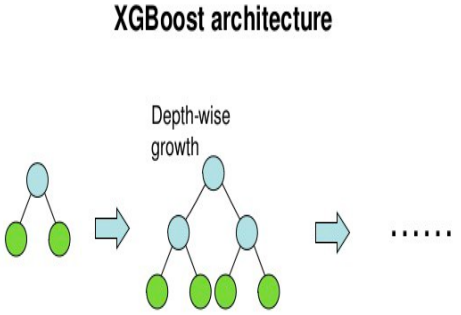

Fig: XGBoost working methodology

Dataset Description:

Having stored that dataset mostly with filename heart.csv into a project folder upon obtaining this from Kaggle. This information was then loaded but also saved to that same dataset's variables using read csv (). They only needed simply glance somewhere at research when it comes to every investigation. Therefore, simply applied using info () function. There really are 13 variables in every and 1 target attribute, as seen. Additionally, there are very few null data, hence there aren't any obvious limitations that worry about.

Features:

(i)Age: Only with chance of contracting cardiac or heart issues roughly double for each decade in one's life, aging becomes the most significant predictor. In puberty, myocardial lipid stripes may start developing. According to estimates, 65 or

older persons make up 82 percent of cardiovascular disease fatalities. This incidence rate also increases over ten years beyond aged 55.

(ii)Sex : shows the person's sex through using following structure:

1 = male ,0=Female

(iii)Chest pain (CP) : indicates the person's kind of chest pains that use the following specifications:

1 = typical angina, 2 = atypical angina

3 = non - anginal pain ,4 = asymptotic

(iv)Blood Pressure(bps) in mm HG : displays a person's mmHg resting blood pressure measurement (unit)

3 is considered average. 6 indicates a defect that has been fixed. Reversible defect = 7

demonstrates whether the person has heart disease: 0. Absence 1,2,3,4 equals present.

(v)Cholesterol in mg: shows the serum cholesterol in milligrams per dl (unit)

(vi)Fasting blood sugar test (fbs): compares a person's fasting blood sugar level to 120 mg/dl. If the fasting blood sugar is greater than 120 mg/dl, then 1 (true) (true) else : 0 (false) (false)

(vii)resting electrocardiographic results: Normal is 0 1 indicates an aberrant ST-T wave. The maximum heart rate that a person can obtain is shown by the formula 2 = left ventricular hypertrophy.

(viii)thalach (Maximum heart rate achieved) : shows the highest heart rate a person has ever reached.

(ix)exang(Exercise induced Angina): Angina induced by exercise 1 = yes 0 = no indicates whether the value is an integer or a float.

(x)oldpeak(ST depression induced by exercise relative to rest) : indicates whether the value is an integer or a float.

(xi)slope (the slope of the peak exercise ST segment): 1.0 =upsloping ,2 = flat when 3 = downsloping.

(xii)ca (Number of major vessels) : The value is displayed as an integer or float

(xiii)thal (Reversible defect): 3 is considered normal. 6 indicates a defect that has been fixed. Reversible defect = 7

(xiv)Target (0,1): demonstrates whether the person has heart disease: 0=Absence and 1equals present.

Results and Analysis:

The study provides an overview of algorithms for heart disease identification utilizing machine learning. Throughout this study, four methods of Machine learning techniques towards prediction of heart disease were examined;

They are SVM, Decision tree, Random Forest, XGBoost. The visualization of every variable with respect to target is done to have an idea on data.Two such examples are shown below:

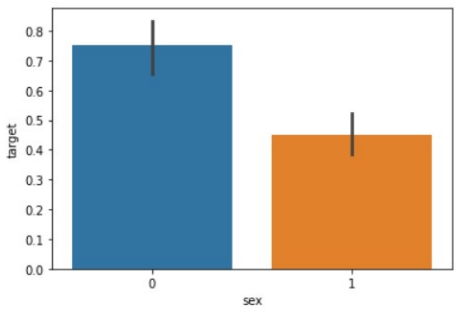

Fig: visualization based on sex vs target

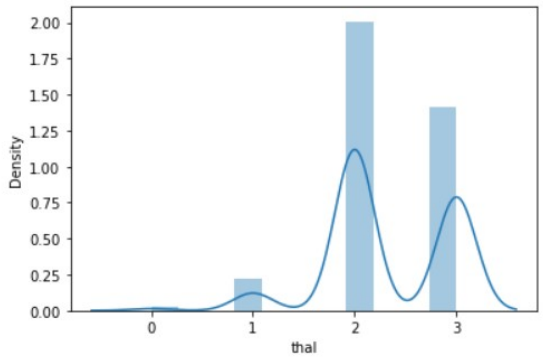

Fig: Density vs Thal

After testing the data, the accuracies of the trained models are as follows:

SVM: 81.97

Decision Tree : 81.97

Random Forest : 90.16

XgBoost: 78.69

```
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 90.16 %
The accuracy score achieved using XGBoost is: 78.69 %
```

So, we suggest using random forest in heart disease prediction is best. By expanding the heart disease dataset's properties but also providing it a little more user-interactive, this study can indeed be enhanced throughout the future. It might also be done using a smartphone app with less intricacy as well as computational cost. When connecting the software towards the hospital's information, we would modify it.

References:

[1] C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", https://www.researchgate.net/publication/323277772_Survey_on_prediction_and_analysis_the_occurrence_of_heart_disease_using_data_mining_techniques, 118(8):165-173 · January 2018.

[2] Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDUL GHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK", https://www.researchgate.net/publication/328630348_PREDICTION_OF_HEART_DISEASE_USING_ARTIFICIAL__NEURAL_NETWORK. October 2018.

[3] Komal Kumar Napa, G.Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", April 2020.

[4] Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach,January 2019.

[5] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open-source solution," Future Science OA, vol. 7, no. 6, Article ID FSO698, 2021.

[6] 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.

[8] 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.

[9] R.Thanigaivel, Dr. K.Ramesh Kumar, "Review on Heart Disease Prediction System using Data Mining Techniques", Asian Journal of Computer Science and Technology (AJCST).

[10] http://www.anderson.ucla.edu

[11] Shadab Adam Pattekari and Asma Parveen, "Prediction system for heart disease using naïve bayes ", International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624. Vol 3,Issue 3, 2012, pp 290-294. 2012.

[12] Carlos Ordonez, Edward Omiecinski, Mining Constrained Association Rules to Predict Heart Disease, IEEE. Published in International Conference on Data Mining (ICDM), p. 433-440, 2001.

[13] Ms. Ishtake S.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research,2013.

[14] Rishi Dubey , Santosh chandrakar "Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground" INDIAN JOURNAL OF APPLIED RESEARCH August2015.

[15] G. Purusothaman , P. Krishnakumari ," A Survey of Data Mining Techniques on Risk Prediction: Heart Disease" , Indian Journal of Science and Technology , June 2015.

[16] Mrs.G.Subbalakshmi , Mr. K. Ramesh ,Mr. M. Chinna Rao , "Decision Support in Heart Disease Prediction System using Naïve Bayes" G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)2011. [17] Bala Sundar V, "Development of Data Clustering Algorithm for predicting Heart", IJCA, Vol 48(7), June 2012, pp 8-13.

[17]Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[18] Buechler K F & McPherson P H (1999). U.S. Patent No. 5,947,124. Washington, DC: U.S. Patent and Trademark Office.

[19] Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[20] Worthen W J, Evans S M, Winter S C & Balding D (2002). U.S. Patent No. 6,432, 124. Washington, DC: U.S. Patent and Trademark Office.