Early Detection of Cardiac Disease Using Machine Learning

Paras Praful Chavda
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, University of Mumbai
paras.chavda@somaiya.edu

Yash Manoj Pithadia
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, University of Mumbai
yash.pithadia@somaiya.edu

Harsh Hirenbhai Bhavsar
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, University of Mumbai
h.bhavsar@somaiya.edu

Radhika Kotecha
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, University of Mumbai
radhika.kotecha@somaiya.edu

Abstract— Heart related diseases are primarily the main reason of death throughout the world and due to which a large number of casualties are arising in countries with low and middle income like India. A large amount of data is continuously generated by medical practitioners. The data generated can be used for the early detection of cardiac diseases, which can effectively support to reduce the occurrence of various heart related diseases. The decision prediction can be effectively done by enhancing the knowledge identification required to discover patterns that were not formerly known. Efficient prediction can be done by accessing the data accumulated from health care companies and industries and find the hidden patterns. The proposed work uses a machine learning algorithm on cardiac-related data and attempts to detect the possibility of cardiac diseases prior to suffering from serious issues. Implementation results demonstrated in the paper show the effectiveness of the proposed approach in early prediction of cardiac diseases.

Keywords— Cardiac disease, Classification, Decision tree, Machine learning, IoT.

I. INTRODUCTION

The motivation of this work comes from the fact that Cardiovascular Disease (CVD) is the number one cause of death globally, more people die annually from CVDs than from any other cause. Another motive to develop the proposed system was that, in the current system a person who intends to check his cardiac health needs to visit the doctor regularly and get all the medical tests done as per advise of the doctor and show the respective reports to him. Depending on the status of the reports, the doctor examines the patient and if found problematic, the patient is treated accordingly. Apart from this there is no way by which an early detection of a cardiac disease can be done.

Especially in India, the number if people died due to heart disease is much more than the casualties caused in the European countries [1]. Deaths due to cardiac disease in has increased continuously in India, it rose by around 41 percent from 155.7 to 209.1 deaths per one lakh population [2]. In children, heart failure can present as respiratory distress, easy fatigability, poor tolerance to exercise, etc.

The pervasiveness of cardiac disease and stroke has rapidly increased by over 50% from 1990 to 2016 in India, with an increase observed in every state. As a result, it can be stated that the number of total deaths and disease burden in the country has almost doubled in the past 25 years [3]. The ratio of deaths and disability because of heart disease was significantly higher in men compared to that in women. There was a rapid increase in deaths due to cardiovascular diseases which was accounted to rise from 13 lakh in the year 1900 to 28 lakh in the year 2016 [3]. The World Health Organization (WHO) has estimated that, with the current burden of CVD, India would lose $237 billion from the loss of productivity and spending on health care over a 10-year period (2005–2015) [4]. Some of the major causes that stimulate the probability of cardiac disease includes: harmful use of alcohol, active smoking raised blood pressure, high cholesterol level excessive presence of fats, lack of physical exercise, high sugar level and an unhealthy diet. The amount of deaths due to cardiac disease was the highest in states like Kerala, Punjab and Tamil Nadu, followed by Andhra Pradesh, Himachal Pradesh, Maharashtra, Goa, and West Bengal.

More than half of the total deaths due to any cardiovascular disease in India in 2016 were in people younger than the age of 70 years. This proportion of death was the highest in less developed states compared with developed and underdevelopment states, which are a major cause for concern with respect to the challenges, posed to the health systems. As a result, reducing premature deaths from cardiovascular diseases in the economically productive age groups requires urgent action across all states of India.

With the increasing number of population across the world and with recent changes of human life styles, there is increasingly higher numbers of individuals with complex medical conditions. Thus, there is an increasingly need for health care systems that can assist with these challenges. Recently, there has been growing attention to the advances within the areas of electronic and medicine engineering and also the nice applications that these technologies offers primarily for health diagnosis monitoring & analysis. Sensors in conjunction with artificial intelligence techniques can be effective for prediction & diagnosis of people with heart diseases with improving lives of people [3].

This paper is organized as follows: In Sect. II, a thorough analysis of the project scope is being done by referring various related popular journal papers. In Sect. III, with the help of statistical analysis and researched papers, a detailed statement of the problem is defined. Sect. IV, depicts the passage from the initial phase of the project to the final phase is described in detail with the help of the modular view of the proposed system. Sect. V, describes the technologies used in the proposed system along with the dataset details. In Sect. VI, model is shown along with the results that have been obtained during the model implementation with the help of respective figures. Sect. VII, concludes the work done in the project. Sect. VIII, highlights directions for future research.

II. RELATED WORK

In [5], the Adaptive Network based Fuzzy Inference System and Linear Discriminant Analysis (ANFIS-LDA) model combined a fuzzy inference method and LDA to predict Coronary Heart Disease (CHD), thus increasing the specificity, sensitivity, and accuracy of the model. In order to increase the prediction accuracy, 625 rules were created by using sample medical data. Further, the classification performance was improved using an ANFIS and LDA. The results of the proposed model showed higher accuracy than those of the existing models. Thus, the proposed model can be used for the prevention of CHD in the general public.

In [6], a mobile monitoring solution is proposed that addresses these challenges and incorporating some smart features to encounter the energy insufficiency of mobile devices and network interruption. The authors of the paper have developed a formal model that evaluated the best execution decision considering online, offline, or combined processing. The model used Dynamic Programming (DP) to determine the best execution path that guarantees an optimum execution time given the resources constraints mentioned above. The paper evaluated the applicability of our solution using electrocardiograph dataset, and proposed paper evaluated the key monitoring processes including preprocessing, feature extraction, and classification.

In [7], the target is to determine the aspects of use of healthcare data which can come to the aid of people by machine learning methods and data mining procedures. The primary objective is to suggest an automated system for diagnosing heart diseases by taking into account earlier information and data. To find out few vital and basic inquiries related with healthcare organization data mining methods are used. For the prediction of heart disease classification models of data mining like Decision tree, Neural networks and Naive Bayes classifier are applied.

In [8], data mining plays an inevitable role in the prediction process of many chronic diseases including deadly heart related diseases. The proposed study examined and revealed the results of applying both k Nearest Neighbor (kNN) and random forests to the Framingham scoring model designed for early risk prediction of Hard Coronary Heart Disease (HCHD). The results reflected that the accuracy of (kNN) was higher compared to random forests in identifying the risk classes among the test dataset compared to the training one. However, the accuracy rate is still to be improved. The authors of the paper suggest using the enhanced kNN approach to enhance the classifier's performance as future scope of work.

In [9], analysis was performed using Classification And Regression Tree (CART) algorithm that yielded ~80% accuracy which was analyzed comparatively well to the performances of different classifiers like Support Vector Machine or Artificial Immune Recognition System (AIRS). From our CART graphical model, it was found that the basis of the tree was feature fifteen, or the center rate in range of beats per minute. As the root of the tree is recommended to be a determinant feature of the info, this is smart as a result of expected traditional or abnormal heart rates to be strongly correlated with arrhythmia. Looking at the top nodes of the tree, it was noticed that these nodes were related to a multiple of different features, but mainly features in the 200 range i.e. channel values.

In [10], before feature selection, Naive Bayes achieves lower cross validation error than SVM. While once feature choice, SVM achieves lower cross validation error than Naive Bayes. The proposed system thinks that the matter could belong the shortage of enough coaching examples (475) and excessive quantity of features (274). Just as it will be seen in drawback, set a pair of once each Naive mathematician and SVM are used to classify spam and non-spam, SVM formula works higher if there are more training examples. However, during this drawback, solely 475 coaching examples are accessible.

To solve the open issues in existing system as well as to have an efficient method of early detection of cardiac disease, the work proposes a Machine Learning based approach in the next section.

III. PROBLEM DEFINITION

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. As per the statistical analysis the risk factors associated with heart disease are identified to be age, blood pressure, smoking habit, total cholesterol, diabetes, hypertension, family history of heart disease obesity, and lack of physical activity. Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Genetic algorithm, Neural network Naïve Bayes, classification via clustering, Decision Tree and direct kernel self-organizing map are some techniques used in the diagnosis of heart disease.

The work presented in this paper is to develop a heart disease early prediction system using machine learning. The solution uses the real-time data gathered using the human body parameters (heart rate and peripheral capillary oxygen saturation (SpO2)) obtained through sensors as well as the parameters of patient like age, blood pressure, cholesterol, etc. entered manually by the user in the mobile application. The heart rate and SPO2 are entered by user by the output LCD from sensors. These parameters are used by machine learning technique Decision tree to identify the risk of cardiac disease. The output shows whether a person is subject to risk of cardiac disease as well as the reasons for the same.

IV. PROPOSED APPROACH

This proposed paper used Decision trees as they implicitly perform feature selection & can tackle nonlinear relationships between parameters. A tree can be "learned" by simple splitting the source set into various subsets based on the attributes and t recursion is completed when splitting no longer adds value to the predictions. Gini and entropy are used to decide the information gained after dataset is split. The proposed approach is shown in Fig. 1.

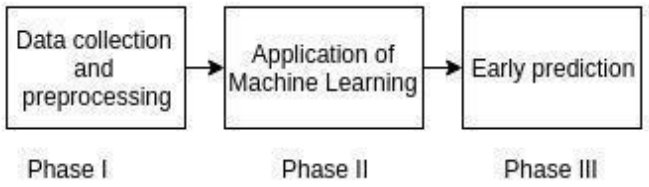

Fig. 1: Modular view of proposed system

In Phase I, the proposed system performs the task of data collection and preprocessing. Phase I exhibits actual implementation of Arduino UNO along with MAX30100 sensor which is used to obtain real time data which comprises of the parameters like heart rate and SpO2. In order to obtain the data, user needs to place the index finger over the MAX30100 module which then captures the data with the help two LEDs, one emitting a red light, another emitting infrared light. The data captured by the sensor can be viewed using a serial monitor present in the Arduino IDE. Collected data is then passed into Phase II to train the machine in order to obtain predicted results.

In Phase II, the proposed system uses python, which is widely used for machine learning as it has simple syntax unlike C++ and java. Libraries like Scikit-learn, Pandas, Numpy gives a great hand to build a machine learning algorithms. Jupyter notebook on the other end gives interactive cell structure for data cleaning, data visualization to machine learning. Data preprocessing is done on the Jupyter notebook also the functions like cleaning the data, finding redundant entries, removing null values can be done easily with the help of Pandas and Jupyter notebook. This preprocessed data is now ready to train a machine, to give the best accuracy on given dataset. The advantage of using the decision tree is that it gives out the feature that have caused the disease to occur. Information gain on all of the features is calculated and the highest gain of information gives the first split of the decision tree, based on that feature. To reduce the effect of bias resulting from the use of information gain, a variant known as the gain ratio is used. The information gain may have taste toward several outcomes. Gain magnitude relation adjusts the data gain for every attribute to permit for the uniformity and breadth of the attribute values. The algorithm gives an accuracy of 71% on 270 entries of data with 14 features. The input data which come from the user then get passed to the machine which gives the predicted results and also the main advantage of using the decision tree is that it gives out the feature that have caused the disease to occur.

In Phase III, with the help of a mobile application the data entered by the user as well as the real time data captured from Phase I is then passed to a trained machine in the form of a data frame. The user enters the data based on the data gained from the display and from the medical reports. Using the mobile application, the user enters the data into the algorithm. The algorithm is present in the back-end of the application. The Python libraries are used to integrate the machine learning algorithm with the mobile application. Entering data to the application is effortless. The user interface is simple to understand and hence it makes the application easy to use for general public. Depending on the data, machine predicts the result which is displayed on the output user interface. Further users are notified for an early diagnosis of heart disease by using alerts.

The system diagram is depicted in Fig.2

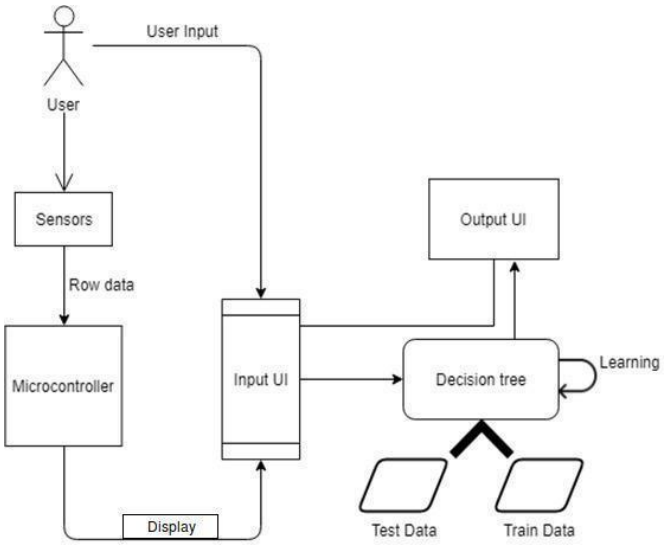

Fig. 2: Proposed System diagram

V. IMPLEMENTATION DETAILS

A. Technology used

Considering the project as a system containing several parts, the proposed system can be classified into two main units, the hardware and the software units:

Hardware of this project comprises of integration of three units namely, the data acquisition, the microcontroller communication unit and alarm unit.

Data Acquisition Unit: This unit is solely responsible to obtain patient's vital parameters utilizing sensors. Sensors can be defined as devices which are used to detect any environmental as well as scientific variations. These sensors can be broadly classified in two types which are Optical sensors and Solid state sensors. The sensors used here are used to detect the heart-rate and the SPO2 measurements.

MAX30100 Sensor: The MAX30100 sensor is an integrated heart-rate and pulse oximetry monitor sensor which comprises of two LEDs, a photo detector, optimized optics, and low-noise analog signal processing which effectively detects pulse oximetry and heart-rate signals. The MAX30100 sensor can operate on the power supply that ranges from 1.8V and 3.3V and can be powered down through software with negligible standby current, which allows the power supply to remain connected at all times.

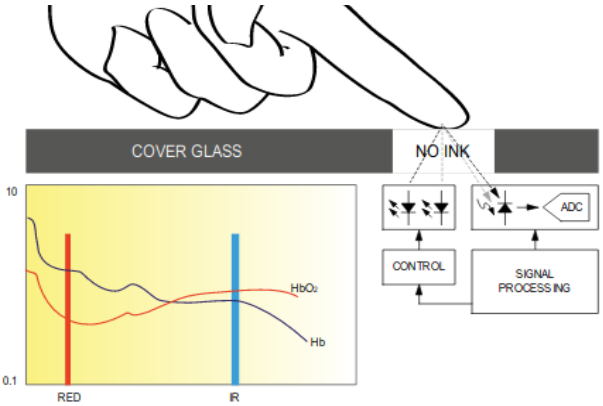

Fig. 3: MAX30100 diagram [28]

In Fig. 3 sensor MAX30100 is shown along with it's working.

Microcontroller -- Arduino UNO: A microcontroller unit is used to filter the received signals from a sensor, apply calculations on them and compose them for transmission to the next unit. In this project, the microcontroller unit chosen is a UNO Arduino board which is based on ATMEGA328 controller.

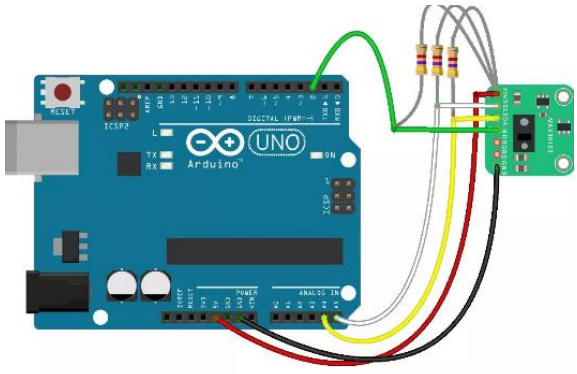

Fig. 4: Arduino interface with MAX30100

In Fig. 4 Arduino UNO is shown interfaced with MAX30100 sensors with the help of three 4.7 ohms resistors which acts as a pull up resistors to provide adequate power supply to the sensor.

Several software tools were used throughout the entire development procedures of this project. In order to program the Arduino board and develop a prediction system besides developing the android application which will detect and alert when suspecting a heart disease.

Arduino IDE: Is the required software environment to program the Arduino by writing a code and embedding it to the Arduino Uno. It conjointly outputs the results for analysis utilizing each serial monitor and serial plotter. The version 1.8.3 is used here which supports both serial monitor to print the heart rate as well as the SpO2 level detected by the sensor.

Python Libraries: Panda and Numpy**,** Numpy is numeric python, widely used for its fast mathematical computation on arrays and matrices. Pandas is used for data analytics. Pandas works on the 2d table object known as data frame. Before using Scikit-learn, put SciPy. The module, as such, provides learning algorithms and is thought as Scikit-learn.

Apache Cordova: Cordova is a platform independent mobile application development framework. It is used to build mobile applications using technologies such as HTML, CSS, JavaScript and many other technologies used to develop websites.

B. Dataset Details

Details of datasets used in the project are mentioned in Table 1.

Table 1: Dataset Details

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest Pain Type 1 = typical angina 2 = atypical angina 3 = non angina pain 4 = asymptomatic
Trestbps	Continuous	Rest blood pressure(in mmHg)
Chol	Continuous	Serum Cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or define left Ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0-3
Thal	Discrete	3 = normal 6 = fixed defect 7 = reversible defect
Diagnosis	Discrete	0 = healthy 1 = patient who is subject to possible heart disease

In Table 1 there are 14 attributes used in the proposed system, which includes 8 symbolic features and 6 numeric features. These features are given as follows: age (defines age in years), sex (explains whether the user is male or female), Chest pain type classified as (typical angina, atypical angina or non-anginal pain, asymptomatic), Trestbps (signifies resting blood pressure in mmHg), cholesterol (signifies serum cholesterol in mg/dl), fasting blood sugar < 120 mg/dl (shows whether it's true or false), resting electrocardiogram (signifies normal, having ST-T wave abnormality which shows probable or definite left ventricular hypertrophy by Estes' criteria), maximum heart rate, exercise induced angina (true or false), oldpeak (ST depression induced by exercise relative to rest), slope (whether up, flat or down), number of vessels colored by fluoroscopy (in the range of 0-3), thal (normal, fixed defect, reversible defect), and class (healthy, with heart disease).

VI. RESULTS AND DISCUSSION

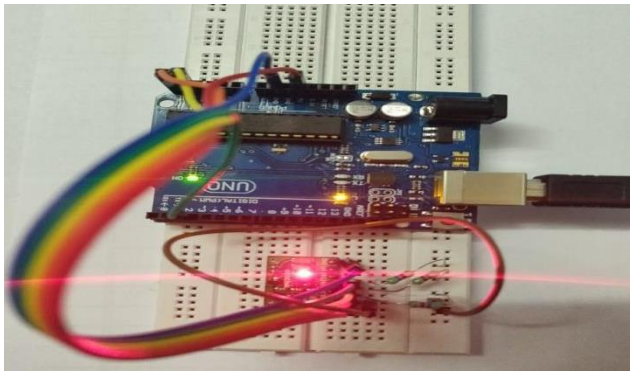

Fig. 5: Actual IOT implementation

Fig. 5 depicts the implementation of Arduino along with Max30100 sensor which forms IOT module that is used to accumulate user data.

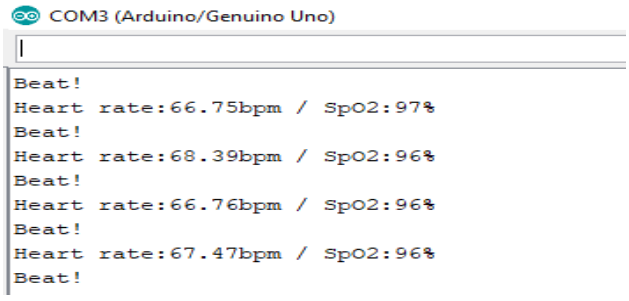

Fig. 6: Output obtained from serial monitor

Fig. 6 depicts the output obtained from the Arduino microcontroller which comprises of parameters such as heart rate and SPO2.

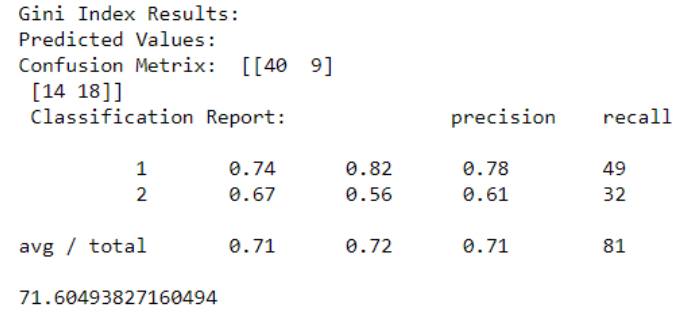

Fig. 7: Decision tree accuracy

Fig. 7 shows the results of algorithm accuracy with confusion matrix.

Fig. 8: Oxywatch measurements

Fig. 8 depicts the actual output of Oxywatch device used by the doctors to calculate patients heart rate and Spo2.

VI. CONCLUSION AND FUTURE WORK

Heart diseases when aggravated spiral way beyond control. Diseases related to heart are much more complex than other diseases and a large amount of people have lost their lives due to such diseases. In small period of time the patients have to face serious consequences if the early symptoms of heart diseases are ignored. Sedentary lifestyle and excessive stress in today's world have worsened the situation. Thus, the early detection of such diseases assist in keeping them in under control. However, daily exercise and getting rid of unhealthy habits is advised. The odds of getting stroke and heart diseases grows due to consumption of tobacco and unhealthy diets. Hence, the work proposes a mobile application that uses human body parameters retrieved from sensors as well as entered by user and applies decision tree classification algorithm in machine learning to predict risk of cardiac disease.

The work can be extended to use real data of patients from hospitals across the nation. The data will have to be pre-processed to make it suitable for applying machine learning. Further, an intelligent system may be developed as a future work that can lead to selection of proper treatment methods for a patient diagnosed with heart disease. Developing a model which trains the machine using the parameter of Electrocardiogram would also increase the accuracy of prediction of an early heart disease.

VII. REFERENCES

[1] S. Reddy K, B. Shah, C. Varghese, A. Ramadoss, "Responding to the threat of chronic diseases in India", The Lancet, 2005.
[2] V. Fuster, "Cardiac statistics", Journal of the American College of Cardiology, Elsevier, 2016.
[3] S. Spencer, R. Horton, "The Lancet Cardiology Collection", The Lancet, 2012.

[4] D. Prabhakaran, S. Yusuf, S. Mehta, J. Pogue, A. Avezum, A. Budaj, L. Cerumzynski, M. Flather, K. Fox, D. Hunt, L. Lisheng, M. Keltai, A. Parkhomenko, P. Pais, S. Reddy, M. Ruda, T. Hiquing, Z. Jun, "Two-year outcomes in patients admitted with non-ST elevation acute coronary syndrome: results of the OASIS registry 1 and 2", Indian Heart Journal, 2005.
[5] J. Yang, J. Kim, U. Kang, Y. Lee, "Coronary heart disease optimization system on adaptive-network-based fuzzy inference system and linear discriminant analysis (ANFIS-LDA)", Personal and Ubiquitous Computing, 2013.
[6] M. Masud, M. Serhani, A. Navaz, "Resource-Aware Mobile-Based Health Monitoring", IEEE Journal of Biomedical and Health Informatics, 2017.
[7] M. Gandhi, S. N. Singh, "Predictions in heart disease using techniques of data mining", International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015.
[8] H. Elsayed, L. Syed, "An Automatic Early Risk Classification of Hard Coronary Heart Diseases using Framingham Scoring Model",The Second International Conference on Internet of things, Data and Cloud Computing, 2017.
[9] P. Rajbhandary, B. Zhou, "Detecting Heart Abnormality using ECG with CART", unpublished, 2019.
[10] L. Chen, Q. Cao, S. Li, X. Ju, "Predicting Heart Attacks", International Journal of Computer Applications, 2018.
[11] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", in proceedings of the international workshop on Healthcare information and knowledge management, pp.17-24, 2016.
[12] M. Shouman, T. Turner, R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients", in proceedings of the Ninth Australasian Data Mining Conference, 2011.
[13] C. Ordonez, "Association rule Discovery with Train and Test approach for Heart Disease Prediction", Journal of Intelligent Data Analysis, 2011.
[14] T. Tavares, A. Oliveira, G. Cabral, S. Mattos, R. Grigorio, "Processing unbalanced data using weighted support vector machine for prediction of heart disease in children", in International Joint Conference on Neural Networks (IJCNN), 2013.
[15] A. Mdhaffar, I. Rodriguez, K. Charfi, L. Abid, B. Freisleben, "Complex event processing for heart failure prediction", IEEE Transactions on NanoBioscience, 2017.
[16] F. Miao, Y. Cai, Y. Zhang, X. Fan, Y. Li, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest ", in IEEE Access, 2018.
[17] E. Loukis, M. Maragoudakis, "Heart murmurs identification using random forests in assistive environments", 3rd International Conference on Pervasive Technologies Related to Assistive Environments, 2010.
[18] N. Allahverdi, S. Torun, I. Saritas, "Design of a fuzzy expert system for determining of coronary heart disease risk", in Proceedings of International Conference on Computer Systems and Technologies, CompSysTech, 2007.
[19] N. Khateeb, M. Usman, S. Zulfikar, "Efficient heart disease prediction system using K-nearest neighbor classification technique" in proceedings of the International Conference on Big Data and Internet of Thing, pp.406-409, 2017.
[20] R. Bialy, M. Salama, "An ensemble model for heart disease data sets: a generalized model" in proceedings of the 10th International Conference on Informatics and Systems, 2016.
[21] J. Ghosh, M. Valtorta, "Building a bayesian Network model of Heart disease" in proceedings of the 38th annual on Southeast regional conference, pp.239-240, 2000.
[22] H. Kahtan, K. Zamli, "Heart disease diagnosis system using fuzzy logic" in proceedings of the 2018 7th International Conference on Software and Computer Applications, pp.297-301, 2018.
[23] S. Fuicu, A. Avramescu, "Real time E-health system for continuous care" in proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp.436-439, 2014.
[25] Pulse Oximeter and Heart-Rate Sensor IC for Wearable Health, MAX30100, http://www.datasheets.maxim integrated.com.