Logistic Regression and KNN Algorithm Experimental Diagnosis to Reduce the Impact of Cardiac Arrest

R.Kannan [1], V.Vasanthi [2]

[1] *Department of Computer Science*, Rathinam College of Arts and Science, Coimbatore, India. e-mail: dschennai@outlook.com
[2] *Department of ICT*, Sri Krishna Adithya College of Arts and Science, Coimbatore, India. e-mail: vasanthiv@skacas.ac.in

***Abstract* - Occurrences of many current diseases can rapidly increase in comparison to prehistoric periods. These diseases are generally hazardous to human lives, one of which is notably the cardiovascular disease. Many medical studies have been conducted with developing related technologies, but despite progress early detection and remedies for cardiovascular diseases remain challenging to the medical sector and physicians. Nonetheless, by using machine learning techniques physicians would be able to accurately predict and diagnose symptoms of pending cardiovascular disease. This paper describes our study, wherein the use of various patient data sets and three key machine learning algorithms help medical professionals appraise and visualize the most precise information relating to heart patients, and also provides vital pre-diagnostic data, as well as information about nearby hospitals, all of which can help to save the patient's life.**

Keywords - component; logistic regression, K-Nearest-Neighbour, KNN algorithm, ROC Curve, cardiac diseases, experimental

I. INTRODUCTION

Despite scientific studies and technological advances, some diseases are incurable even if we detect the inception of the diseases. Notably, the detection of cardiovascular diseases is still a challenge to the health sector and medical professionals. Hospitalization costs and the costs of medical care are a major concern to both the caretaker and patient. Various areas of research show that cardiovascular mortality is high in developing countries. This increase is attributed to the lack of awareness by the public, the consequences of a sedentary lifestyle, improper habits, genetic factors, diabetes mellitus, high blood pressure (HBP), and birth complications. Owing to lack of awareness about early detection and first aid for cardiac arrests, many patients lose their lives before reaching the hospital [8].

Unfortunately, many people are unaware about cardiopulmonary resuscitation (CPR) as an emergency procedure that can be administered during cardiac arrest. Furthermore, research shows that 50% of the patient's life could be saved if proper first aid is provided at the time of cardiac arrest. It is in these kinds of situations that machine learning would be beneficial to the medical professionals, as it helps to improve the chances of saving the patient's life.

In recent times, researchers have started using machine learning techniques to help diagnose cardiovascular disease [1]. The machine learning techniques and "big data" play a tremendous role in the healthcare sectors and medical industries. Machine learning identifies the risk patterns in high volumes of data and provides conclusive solutions to the medical problem with the help of these existing data. By using the vast and readily available existing datasets about the patient, a precise prediction and diagnosis of the symptoms of cardiovascular disease, together with the severity of a heart attack, can be automatically predicted. Visualization and monitoring of the patient's data in real time can also be performed.

In this study, we have utilized vast amounts of data from four different datasets, which informatively helps the medical professional to easily and accurately diagnose signs of a pending heart attack.

The following techniques are used: (i) high volumes of data processed with the high-performance Apache Spark as a unified analytics engine for large-scale data processing (big data), (ii) a logistic regression algorithm to help predict the signs of a pending heart attack, (iii) a Receiver Operating Characteristics (**ROC**) curve in the logistic regression technique to improve the prediction performance, (iv) the K-Nearest-Neighbour (**KNN**) method to easily find the nearest hospitals, and (v) the R programming language and big data to monitor and provide on-time visualization of the patient's data reports by medical professionals. Using the above techniques, the medical professional can provide early diagnosis and recommend hospitalization and treatment in a shorter time than normal to save the patient's life.

II. LITERATURE REVIEW

In recent years' various machine learning techniques have been employed to predict heart disease by the researchers, but the obtained accuracy is not same. Some of the techniques and datasets used for prediction as well as results have been discussed here.

Liangqing Zhang, Cuirong Yu ROC Technique can be performed to keep patients under continuous supervision by using a medical monitoring system for heart diseases. 29 features of attributes have been extracted after Different sessions of experiments were conducted with the psychological data collection of seven DAYS and 30 days. In the final ROC curve of prediction using 30 days and seven days reached 67.6% and 79.4%.

Mamta Sharma, Farheen Khan and Vishnupriya Ravichandran showed that comparison of the results of artificial Intelligence with data mining techniques such as decision tree and naïve Bayes for heart diseases prediction have been done. In this research work, Heart disease prediction system was developed by using 13 attributes and 15 attributes. The Artificial Neural network gives 100% prediction accuracy by using 15 attributes, whereas data mining techniques or not so accurate.

Vladimir S. Kublanov,Anton Yu. Dolganov Have presented an efficient approach for arterial hypertension diagnostics for analysis of cardiac activity by using various machine learning technique. This study was conducted with 30 healthy volunteers and 41 patients suffering from arterial hypertension. In the final linear discriminate analysis, 91.33% results have been achieved and proved to be clear as compared to other machine learning technique such as kNN, SVM, DT and NB.

After reviewing above literature the researchers have come to a conclusion that they got different prediction accuracy results. Whereas roc method employed in this research led to 93% prediction accuracy. In addition, other techniques have been used to alert and visualize to health professional in order to save the patients from the risk of heart attack.

III. Materials & Methods

In this section, technical devices such as smart phones and wireless heart monitor systems are employed. Additionally, the American Heart Association (AHA) datasets are compared and analyzed with real time datasets using Spark Big Data. Finally, the datasets are integrated with the machine learning techniques to predict, visualize, alert, and diagnose the heart disease.

A. Wireless Heart Rate Monitoring System and Smart Phone

A wide range of heart monitoring sensors and smart phones are available in the market. To overcome power consumption issues with the use of such devices in real time, we have selected Wahoo X, as this specific device consumes low energy, has high usability and provides good performance. This wearable sensor extracts and records heart rate 6 times per minute.

B. Heart Rhythm Dataset (HRD)

The Wahoo X wearable sensor extracts and records heart rate and other physiological data 6 times per minute, as shown in Table I.

TABLE I. HEART RHYTHM DATASET

Field Name	Description
Patient ID	Patients ID requires to identify and track the heart diseases by hospitals
HR	heart rate [bpm] (numeric)
Date Time	date and time (format "YYYY-MM-DD HH:MI:SS")
GPS Landmark	latitude and longitude (GPS X & Y coordinates)
Steps	steps (numeric)
GSR	galvanic skin response (numeric)
Calories	burned calories (numeric)
Temp	skin temperature [°F] (numeric)

C. American Heart Association (AHA) Dataset

We have selected the AHA dataset for comparison with the patient's on-time Heart Rhythm dataset to provide alarms with regards to any detected heart abnormalities [18]. In addition, we have discussed and confirmed the classification of this dataset with heart specialists (shown in Table II below).

TABLE. II AHA PREDICTED HEART DISEASES DATASET

Sinus Rhythm Type	Threshold value of HR, BP and hotness
Normal	60<=HR<=100 (Beats /min), BP=100-140/60-80 mmHg, & Hotness =36.5-37.5 o C
Bradycardia	HR<=60 (Beats /min)
Tachycardia	HR>=100 (Beats /min)
Hypertension Stage 1)	BP=Sys/Dys>=140/90 mmHg
Hypertension (Stage 2)	BP=Sys/Dys>=150/95 mmHg
Hypotension	BP=Sys/Dys<=00/60 mmHg
Fever	Hotness >=37.8 o C
Hypothermia	Hotness <=35.0 o C

D. Patient's Dataset

This dataset contains 20 attributes of the patient's general details, such as name, contact details, previous diagnostic history, details of the patient's family member's, etc. Moreover, this dataset is utilized to provide diagnosis based on the patient history, allowing the sending of an alert notification when an emergency occurs (shown in Table III below).

TABLE III. PATIENT DATASET

Field Name	Description
Hospital ID	Patient's ID required to identify and track the heart diseases by hospitals
Hospital ID	Patient's registered Hospital ID and valid Hospitals listed in Hospitals dataset
Patient ID	Patient's ID requires to identify and track the heart diseases by hospitals
First Name	Patient First Name
Last Name	Patient Last Name
Mobile No	Patient's Mobile No to track and extract the data from wireless hrms device
Address	Patient's permanent address along with postal code for communication
Date of Birth	Valid date>1900 and <=Today.
Gender	Male/Female/Not known
Height	Patient's Height in cm
Weight	Patient's Weight in kgs
Diagnosis History	The patient's diagnosis existing records
Previous Heart failure	Hypertensive heart disease with heart failure, Ischemic / Dilated cardiomyopathy, cardiomyopathy, cardiomyopathy unspecified, congestive heart failure, left ventricular failure, heart failure unspecified, and not applicable
Smoking status	Never smoked / Ex-smoker / Current smoker
Diabetes Status	Nondiabetic/Diabetics (dietary control)/Diabetic (oral medicine)/Diabetic (insulin)/Insulin plus oral medication
Family Member(s) Names	To communicate with the family members when the patients occurs heart
Family Member(s) Mobile Numbers	To alert & communicate with the family members when the patients occurs heart
Wireless HRMS Status	Yes/No
Wireless HRMS Number	Wireless heart rating monitoring system unique number
Wireless HRMS Name	Wireless heart rating monitoring system Model Name
Wireless HRMS Details	Wireless heart rating monitoring system Model technical specification

E. Hospital Dataset

This dataset includes a total of 12 attributes, such as hospital details, heart specialist details, and many more (shown below in Table 4). This allows emergency personnel to find the closest hospitals, doctors' availability, ambulance availability, and the necessary amenities in case of emergency.

TABLE IV. HOSPITAL DATASET

Field Name	Description
Hospital ID	Registered Unique Hospital ID
Hospital Name	Name of the hospital
Contact Numbers	Hospital Contact numbers (24/7)
Location	Exact Hospital Location
GPS landmark	latitude and longitude (GPS X & Y coordinates)
Address	Hospital Address with postal code
Working Days	Hospital working days in a week
Working Hours	Hospital working hours per day
Specialists Name	Names of the heart specialists
Specialists Mobile Number	To communicate, identify and track the heart patients diseases
Ambulance facility	Yes / No

F. System Architecture

The details of system architecture shows the major components that extract and compare the data required for the prediction method, as shown in Fig. 1. Also detailed are the techniques used for prediction, monitoring, provision of alerts, and data visualization.

First, we use the wearable sensor device Wahoo X to collect a person's physiological data (shown in Table 1). The smartphone extracts the patient's data from the sensor device which is conveyed via the Bluetooth device. Extracted data will be read by the Spark big data engine and then processed for visualization by the healthcare professional by using the approach of Matplot in R programming.

Meanwhile extracted data will be analysed and compared with AHA dataset using the logistic regression method to detect the presence or absence of a heart problem. Logistic regression includes two types of methods, called regression and classification, for evaluating the dataset. It is independent of explanatory variables that conclude an outcome. Meanwhile, logistic regression searches the best fitting method to describe the association between response and predictor variables. However, this method replaces the linear function with sigmoid function and gives two possible outcomes (shown in Fig 2). The logistic regression defined by the following equations.

$$Y = Logistic(c + x_1 * \omega_1 + x_2 * \omega_2 + \cdots + + x_n * \omega_n)$$

$$Y = \frac{1}{1 + e^{-(c + x_1 * \omega_1 + x_2 * \omega_2 + x_3 * \omega_3 + \cdots + x_n * \omega_n)}}$$

Finally, the logistic function with the sigmoid formed by the following equation.

$$Y = \frac{1}{1 + e^{-x}}$$

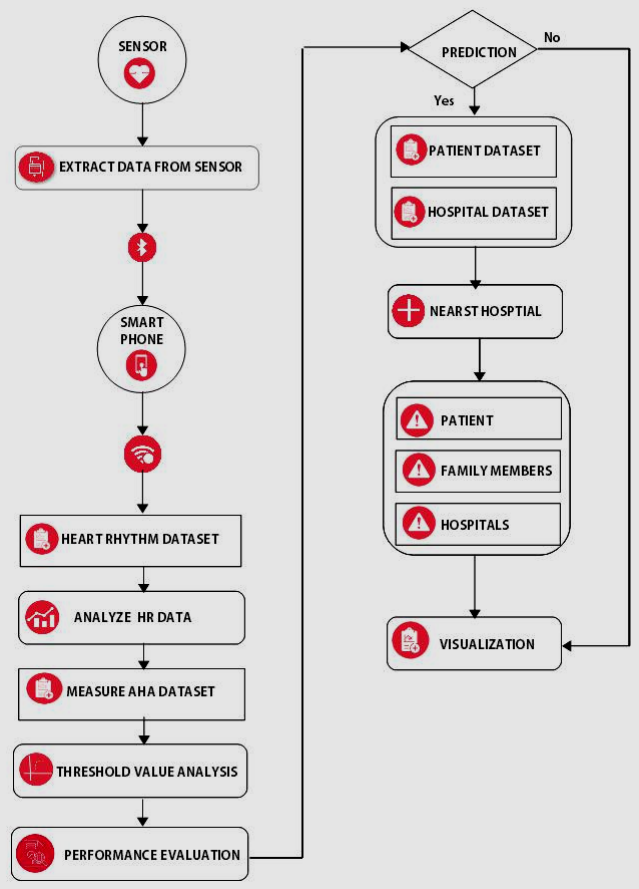

Fig.1 System Architecture

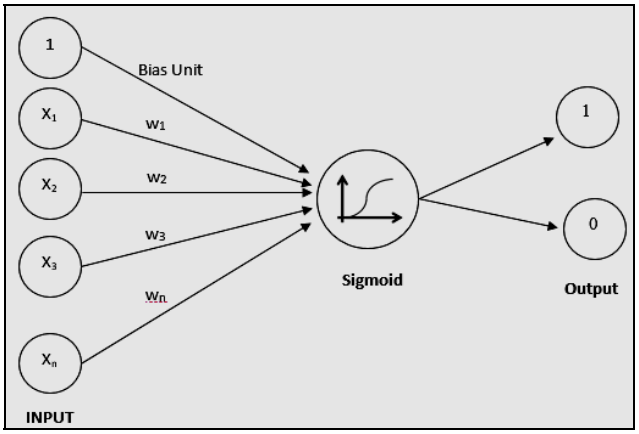

Fig.2 Logistic Regression

In addition, the ROC curve method is used, which improves the performance efficiency by the model-wide evaluation measure from the discombobulation matrix.

The model-wide evaluation measure separates the datasets into two parts as positive and negative and uses the below performance measures to measure the performance of a separate dataset.

- Sensitivity - to measure the positive part of a dataset.
- Specificity - to measure the negative part of a dataset.

The specificity and sensitivity are as provided in the following equation.

$$\text{True Postive Rate (TPR)} = \frac{TP}{TP + FN}$$

$$\text{False Postive Rate (FPR)} = \frac{FP}{FP + TN}$$

$$\text{Postive Likelihood Ratio} = \frac{\text{Sensitivity}}{(1 - \text{Specificity})}$$

$$\text{Negative Likelihood Ratio} = \frac{(1 - \text{Sensitivity})}{\text{Specificity}}$$

The ROC curve plots True Positive Rate vs. False Positive Rate for different classification thresholds [9].

The most positive outcomes of sensitivity performance measures are classified and given by lowering the part of the classification threshold which increases the positives of both True and False. Moreover, for assessing the classification performance (accuracy and average false positives), the threshold values are defined in Table V.

TABLE.V HEART DISEASES THRESHOLD VALUES

Age	18-35	36-64	Above 64
Normal HR	72-75(BPM)	76-79(BPM)	70-73(BPM)
Bradycardia	HR<=55	HR<=60	HR<=65
Tachycardia	HR>=110	HR>=120	HR>=100
Hypertension	BP>=150/100	BP>=145/195	BP>=140/90
Hypertension	BP<85 mmHg	BP<96 mmHg	BP<117 mmHg
Fever	Temp>=37.2ºC	Temp>=37.5º C	Temp>=36.9º C
Hypothermia	Temp<35.5ºC	Temp<35.1ºC	Temp<35.0ºC

If emergencies are indicated after data analysis and prediction, then the structure of this system will find the closest hospitals by using the KNN Euclidean algorithm. KNN is a supervised machine learning algorithm, which is quantified by Euclidean distance functions. Specifically, KNN is used to find the nearest neighbors in the following equations.

$$\text{Euclidean } \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

The Euclidean distance functions can only be valid for continuous variables. Moreover, KNN algorithms take much less time and computes any distances without calculating based on the distance functions.

The Plivo Application Program Interface (API) integrates with R for the support of HTTR. Therefore, the API generates calls and SMS messages automatically to the patient's relatives and the nearest hospitals, according to the

information in the patient's dataset and hospital dataset. Additionally, the approach of using the MYGMAILR package with R automatically generates and sends the notification emails to the nearest hospitals and patient relatives.

IV. Results and Discussion

Our main purpose is to create an intelligent classification method that reveals whether patients have heart disease or not, so that we can predict, alert, and provide visualizations to health professional in order to save the patients from the risk of heart attack. For this purpose, we have used one month of real-time data, which consist of 20 patient's details. The data split is 70% data for training purpose and 30% data for testing purpose using the approach of R programming randomly.

In the next step, the approach of logistic regression method to train and it will be compared with an AHA threshold for classifying that presence or non-presence of heart attack, and categorizing the patients.

After many trials and errors, we have checked the process with the help of AHA dataset and heart specialties and by the tuning parameters and the data tested with the trained model. We obtained only 86% accurate results through this testing process. Moreover, to improve accuracy and performance, the use of the ROC curve method led to a 93% prediction accuracy result, see Fig 3.

Finally, to find the nearest hospital for alerting process while emergency occurs by the approach of KNN Euclidean distance we achieve approximately 83% of the result of alert processing of SMS calls and email (shown in Table VI).

TABLE. VI. PREDICTED TIMES OF ALERT PROCESSING OF SMS CALLS AND EMAIL RESULTS

Internet	Predicted Time
Wi-Fi	11.0 s
4G	11.3 s
3G	13.1 s
2G	19.8 s

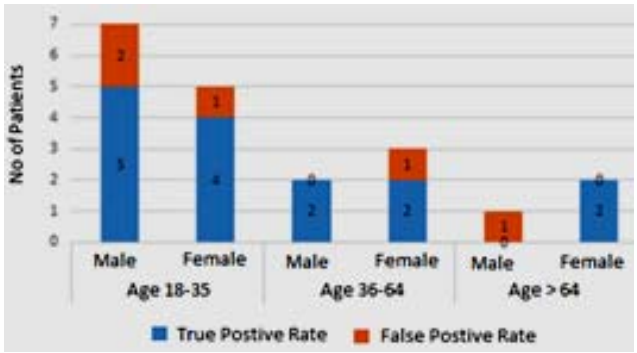

Fig.3.A. Heart Diseases Prediction and Performance Measure

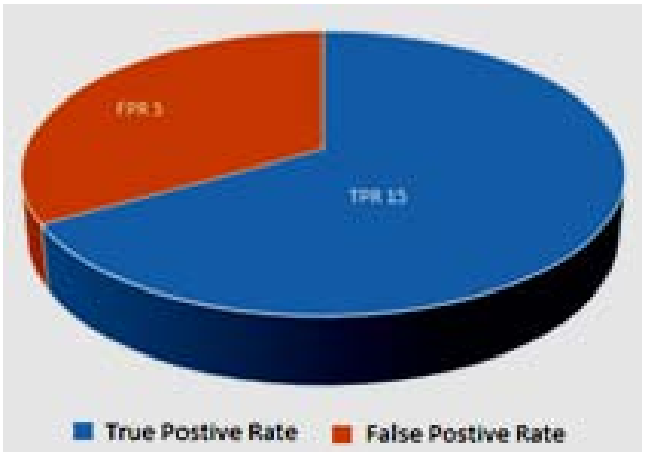

Fig.3.B. Heart Diseases Prediction and Performance Measure

V. Conclusions

This system architecture predicts heart disease very accurately and alerts the condition of the patient in the least possible time, therefore lowering a patient's risk level towards heart disease. Additionally, frequent remote monitoring of the patient's condition by the health professional will reduce the need for frequent hospital visits.

References

[1] Pedro Domingos, "The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World".

[2] Ian H. Witten, Eibe Frank , Mark A. Hall, "Data Mining Practical Machine Learning Tools and Techniques third Edition", Morgan Kaufmann Publishers is an imprint of Elsevier, pp 978-0-12-374856-0.

[3] John Paul Mueller, Luca Massaron, May 2016, "Machine Learning for Dummies".

[4] Karthik Ramasubramanian Abhishek Singh, 2017, "Machine Learning Using R - A Comprehensive Guide to Machine Learning".

[5] Shai Shalev-Shwartz and Shai Ben-David, 2014, "Understanding Machine Learning: From Theory to Algorithms" by Cambridge University Press.

[6] coursera.org, ' Machine Learning', 2017,

[7] https://www.coursera.org/learn/machine-learning.

[8] Roberto Battiti Mauro Brunato, 2014 "The LION Way Machine Learning Plus Intelligent Optimization".

[9] Brendan Phibbs MD, "The Human Heart: A Basic Guide to Heart Disease", Second Edition, University of Arizona College of Medicine, Tucson, Arizona, pp 978-0781767774.

[10] "Introduction to the ROC (Receiver Operating Characteristics) plot", 2018, https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot.

[11] Jason W. Osborne, 2015, "Best Practices in Logistic Regression 1st Edition", University of Louisville, USA, pp 978-1452244792,

[12] Y. Lakshmi Prasad, 2015, "Big Data Analytics Made Easy", pp 978-1-946390-71-4.

[13] Nick Pentreath, 2015, "Machine Learning with Spark", pp 978-1-76326-651-9.

[14] Holden Karau, Matei Zaharia, Andy Konwinski, Patrick Wendell, Matei Zaharia, 2015, "Learning Spark : Lightning-Fast Big Data Analysis", MIT University, USA, pp 978-9351109945.

[15] Daniel T. Larose , Chantal D. Laros, " k - Nearest Neighbor Algorithm", 2014, https://doi.org/10.1002/9781118874059.ch7.

[16] Dr. Saed Sayad, 2018, "An Introduction to Data Science" , http://www.saedsayad.com.

[17] Atmajitsinh Gohil, 2015, "R Data Visualization Cookbook", pp 978-1783989508

[18] Zoiner Tejada, 2017, "Mastering Azure Analytics: Architecting in the Cloud with Azure Data Lake, HDInsight, and Spark", First Edition, pp 978-9352135356.

[19] American Heart Association: https://professional.heart.org/professional/ResearchPrograms/UCM_461443_AHA-Approved-Data-Repositories.jsp.

[20] Wahoo X Heart Rate Monitoring System (HRMS), https://www.wahoofitness.com/devices/heart-rate-monitors.

[21] plivo.com, "Cloud API Platform and Services Provider for Voice Calls and SMS", https://www.plivo.com/docs/getting-started/.

[22] Liangqing Zhang, Cuirong Yu, Chunrong Jin, et al., "A Remote Medical Monitoring System for Heart Failure Prognosis," Mobile Information Systems, vol. 2015, Article ID 406327, 12 pages, 2015.

[23] Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran, "Comparing Data Mining Techniques Used For Heart Disease Prediction," f Engineering and Technology, vol.5, Issue 6, June 2017.

[24] Vladimir S. Kublanov, Anton Yu. Dolganov, David Belo, and Hugo Gamboa, "Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics," Applied Bionics and Biomechanics, vol. 2017, Article ID 5985479, 13 pages, 2017.

[25] Akram SA, Ghaleb S, Hamaid SB, Vasanthi V (2017) Survey study of virtual machine migration techniques in cloud computing. Migration 177(2):19–22.

.