ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Prediction of Heart Diseases (PHDs) based onMulti-Classifiers

Thandra Satish Kumar

Department of CSE**,** Koneru Lakshmaiah Education FoundationVaddeswaram, AP, India

thandravas9989@gmail.com

Vaddi Chakradhara Vasu

Department of CSE, Koneru Lakshmaiah Education FoundationVaddeswaram, AP, India

Chakradhara0050@gmail.com

Tikkana Srinivas

Department of CSE, Koneru Lakshmaiah Education FoundationVaddeswaram, AP, India

srinivastikkana@gmail.com

T.Santhi Sri,

ProfessorDepartment of CSE,Koneru Lakshmaiah Education FoundationVaddeswaram, AP, India sri_santhi2003@yahoo.com

Abstract—

Now a days, heart disease is one among the most complicated issues and globally many people Effected from this disease. A major obstacle to clinical data analytics is the disease's prediction. In time identification of disease is crucial to save patient, so health industry collect data from across world and transforms huge amounts of raw data into useful information. With the help of this useful information, we executed various machine learning algorithms to predict the heart disease. Numerous research have demonstrated that important features are crucial in enhancing the effectiveness of machine learning models. In order to improve patient accuracy and predict patient survival, it is important to identify key traits and efficient data mining approaches. A ML system may detect Cardiovascular disease in its initial days using medical data, lowering fatality rates. Numerous studies have used various ML techniques to recognize Cardiovascular disease or determine the extent of the victim condition. One of the trickiest jobs in medicine is diagnosing and predicting heart illness. Finding the cause of this requires more time, Mostly for doctors and other medical professionals. Massive amounts of unstructured data produced by the healthcare sector are transformed through data mining into information that is helpful for decision- making. Numerous research have demonstrated that important features are crucial in enhancing the effectiveness of machine learning models. In this study, 303 hospitalised patients' heart disease risk was predicted using a dataset. The

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

objective is to identify key characteristics and efficient data mining approaches that can improve the predictability of cardiovascular patients. Six categorization algorithms are used in this research to estimate the customer's mortality: Decision Tree (DT), XGBoost classifier (XGBoostSupport Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (SVM). In this article, we suggest a technique that aims at finding significant features by applying machine learning techniques resulting in improving the performance of CVD.The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 93.41% through the prediction model for heart disease with the logistic regression.

Keywords—Classification; cardiac disease; multiclassifier;heart disease detection; machine learning.

INTRODUCTION

WHO claims that the main cause of death worldwide is heart disease. It is very challenging to determine thecardiovascular disease (CVD) because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors. Sometimes CVD symptoms may change depending on the gender. A male patient, for instance, is more likely to access the US healthcare system. In terms of healthcare services, medication, and lost productivity as a result of death, in the years 2014 and 2015, it cost roughly $219 billion annually. Additionally, early detection might lessen the risk of heart failure, which can cause a person's death.. A number of factors such as blood pressure, cholesterol, creatine, etc., It is challenging to diagnose because these factors affect heart health. The authors of examined many risk factors for heart disease. and identified controllable factors such as alcohol usage, smoking, diabetics, high cholesterol, and limited physical activity.It might contain a few mistakes, and since heart disease is a serious condition, these small mistakes could ultimately result in a death. Expert systems built on machine learning can accurately identify CVD, which lowers the death rate. In order to extract usable information from huge data, data mining is crucial. It is extensively employed in virtually every sphere of human endeavour, including engineering, business, and education. Data mining is the process of examining data to uncover hidden information that can be

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

utilised to make critical decisions in the future. By reducing the inaccuracy in prediction and factual outcomes, a range of machine learning methods have been employed to comprehend the complexity and non-linear interaction between various components.. According to a recent study by the World Heart Federation, CVD is to blame for one in three fatalities. Figures from the World Health Organization (WHO) indicate that by 2030, and over
23.6 million of people could pass away from CVD, primarily due to heart disease and stroke. The existence of normally take, the greatest risk level, the least risky level, and lastly the incorrect diagnosis can all be used to evaluate health status. Due to hereditary considerations, people's socioeconomic and medical condition, environmental circumstances, and individual lifestyle choices, the diagnosis procedure may take longer than anticipated. To get improved health outcomes, contemporary healthcare is generated with anticipating , assessing exposure to prevent sickness before becoming worsens. As a result, a highly accurate method is created that analyses clinical data related to cardiovascular disease (HD) to find the presence of heart problems. Machine Learning (ML) methods were utilised in numerous research to predict CVD from clinical information. Clinical datasets still provide significant challenges because of imbalanced data and complexity. With in healthcare market, databases now include a lot of information about patients with medical reports, and that amount is growing quickly every day. There is a lot of duplication in this uneven raw data. Preprocessing is required to extract crucial characteristics, shorten the runtime of the training algorithm, and improve classification effectiveness. These procedures are now enhanced by the innovative changes in computing capacity and ML regulatory skills, which also create new scientific prospects with in medical sector, especially regarding the previous diagnosis of diseases for conditions like cancer and cardiovascular disease to increase survival rates. Applications for machine learning are numerous, ranging from recognising illness health risks to creating better vehicle safety mechanismsOne of most popular predictive modelling techniques are provided by machine learning to overcome the present limitations. It offers a great deal of potential for processing enormous amounts of data and creating feature sets. It minimises the discrepancy between predicted and actual results to understand intricate and non-linear correlations between attributes. In order to forecast the outcome of an unknown dataset, the computer learns trends from the attributes of the already available dataset. Classification is one of the best machine learning prediction techniques. The supervised machine learning technique that is quite efficient at detecting the disorders when given with the right data. This

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

study's significant element was the development of an understandable healthcare forecasting system for such CVD detection with state-of-the-art ML methods. In this work, different classification algorithm approaches were mastered, including regression models, K-nearest neighbours(KNN), support vector machines, etc.

I. MOTIVATION FOR THE WORK

The primary reason for conducting this study is to propose a model for predicting the development of cardiovascular disease. Additionally, the objective of this study is to determine optimum classification method for detecting cardiac arrest in a patient. Six classification methods, including Support vector machines(SVM), decision trees, random forests, K-nearest neighbours(KNN), and logistic regression, were employed in a comprehensive investigation and evaluation to verify this research at various levels of grading. Even if these machine learning methods are widely utilised, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to assess the six algorithms. This will enable scientists and medical professionals to create a stronger.

II. HARDWARE SPECIFICATIONS

Automated Trading systems may seem quite complicated but only needs a few pieces of gear; all you need is a good computer and an excellent editor. and you're ready to go, not much requirement of extra hardware specifications. Software Specifications

- Jupyter:- Designers will actually carry out the math to distinguish wellsprings of data, carry out checks, and accurately predict the typical results using a Python manager.
- NumPy:- It is essentially a modules, or one might name this a library, that is accessible in Python for data processing right now. It includes 10 solutions for working with C, C++ in addition to a potent dimensionality arrays structure. In linear algebra, it is also very helpful. I'm going to tell you guys that NumPy has Fourier transform and randomly generated skills in addition to being a useful multiple database container for general data. what exactly is multidimensional array now over here this picture actually depicts multidimensional array so we have various elements that are stored in their respective memory locations so we have one two threes in their own memory locations now why is it two dimensional it is two dimensional because it has rows as well as columns so you can see we have three columns

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

and we have four rows available so that is the reason why it becomes a two dimensional array so if I would have had only one row then I would have said that it is a one dimensional array but since it contains rows as well as columns that is it is represented in a matrix form that is why we call this as a two dimensional array so I hope we are clear with what exactly two dimensional arrays

- Matplotlib :-It is a very useful library for just the NumPy technical computing extensions for the Python computer language. It provides a constructive criticism API for inserting charts into proposals that mature into something incredibly useful.
- Pandas:- It is likewise a Python library for the study and surveillance of data. It gives comprehensive operations and steps especially for managing having to contact and arithmetic tabular data.
- Seaborn:- This is a matplotlib-based Python data visualisation package. It offers a sophisticated sketching tool for creating eye-catching and instructive analytical visuals.
- Scikit-Learn/Sk Learn :- The primary computations in this machine learning python package for organisation, recurrence, and clumping involve support vector machines (SVM), incline raising, irregularity woods, and k-Nearest Neighbors.

LITERATURE STUDY

In past years, several studies and analyses have been done in the fields of medical technology and algorithms, resulting in the publication of important publications.

[1] Applying decision tree and hill climbing algorithms, Purushottam suggested a "Efficient Heart Disease Prediction System" in their study. Researchers used the Cleveland dataset, and then before applying classification methods, data was preprocessed. Exponential Learning an opensource data mining programme that fills in the missing values in the data set, provides the basis for the Knowledge Extraction process. A decision tree operates in a top-down fashion. At each level, a node is chosen by a test for every actual node chosen by the hill-climbing algorithm. Confidence are the variables and their corresponding values. Its confidence level is at least 0.25. About 86.7% of the time, the system is accurate.

[2] In his work "Prediction of Heart Disease Using Machine Learning Algorithms," Santhana Krishnan advocated using decision trees and the Naive Bayes method to predict heart disease. The decision tree algorithm builds the tree based on specific circumstances that result in True or False choices. The outcomes of algorithms like SVM and KNN are based on split

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

conditions that can be vertical or horizontal depending on the outcome variable. However, a decision tree is a structure that resembles a tree with a cluster head, branches, and limbs, and it is based on the decisions made in each tree. The value of the attributes in the dataset is also explained by the decision tree. Additionally, researchers used the Cleveland data set. Using some techniques, the data set is divided into 70% training and 30% testing. The accuracy of this method is 91%.The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[3] The research conducted by Sonam Nikhar for her work, "Prediction of Heart Disease Using Machine Learning Algorithms," provides a detailed explanation of the Naive Bayes and decision tree classifiers, which are utilised in particular to predict heart disease. Analysis that considered applying a predictive data mining approach to same dataset determined that Decision Trees have greater precision than Bayesian classifiers.

[4] In their research titled "Prediction of Heart Disease Using Machine Learning," Aditi Gavhane et al. used the multi-layer perceptron neural network approach to train and test their dataset. It will have one input data, one output data, and perhaps more convolutional units throughout this method among the two input and output layers. Every input node is linked to the output layer by hidden layers. Weights chosen at random are assigned to this link. The second input is referred to as bias, and it is given weight according to the needs of the link between the nodes.

[5] In "Heart Disease Prediction Using Effective Machine Learning Techniques," Avinash Golande advocated the use of a few data mining techniques to help clinicians distinguish between different types of heart disease. Nave Bayes, Decision trees, and k-nearest neighbour are commonly used approaches. Packing calculation, Part thickness, consecutive negligible streamlining, neural systems, straight Kernel selfarranging guidance, and SVM are other novel characterization-based procedures that are used (Bolster Vector Machine) .

[6] In his "Machine Learning Techniques for Heart Disease Prediction," Lakshmana Rao argued that there are more risk factors for heart disease. Therefore, it is challenging to identify heart illness. Different neural networks and data mining techniques are utilised to determine the severity of heart disease among patients.

[7] Abhay Kishore suggested "Heart Attack Prediction Using Deep Learning," where a

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

system is utilised to forecast heart attacks through using Deep learning techniques as well as to determine the probability of heart-related infections with in client. To provide the most accurate model with the fewest errors, this model employs deep learning and data mining. This study serves as a reliable benchmark for other cardiac arrest prediction programmes.

[8] To increase accuracy in cardiovascular issues, Senthil Kumar Mohan presented "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." The suggested approach for treating cardiovascular problem combines a linear model and hybrid random forest uses the methods K-Nearest Neighbour(KNN), Linear Regression(LR), Support Vector Machine(SVM), to generate enhanced demonstration scale with an accuracy measure of 88.7%. (HRFLM).

[9] Anjan N. Repaka provided a model that reviewed and contrasted past work and indicated the effectiveness of

forecasting for two classification models. The experimental findings demonstrate that our theory's accuracy in determining the percentage of risk stratification is higher than that of previous models.

[10] Heart Disease Prediction Using Evolutionary Rule Learning," suggested Aakash Chauhan. Electronic records allow for direct data retrieval, reducing the need for manual operations. The range of services is reduced, and it is evident that a high proportion of regulations contribute to the most accurate prognosis of heart disease. On the patient's dataset, pattern matching development connection mining is carried out to produce powerful associations.

PROPOSED SYSTEM

information assets are located, then further chosen, cleansed, and transformed into the required form. To accurately forecast cardiac disease, various classification approaches will be used to data set. The accuracy of several classifiers is compared usingthe efficiency metric.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

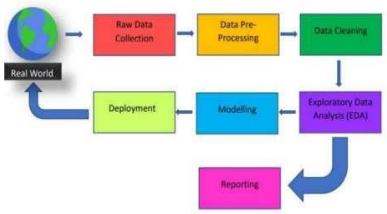


Fig 5.1

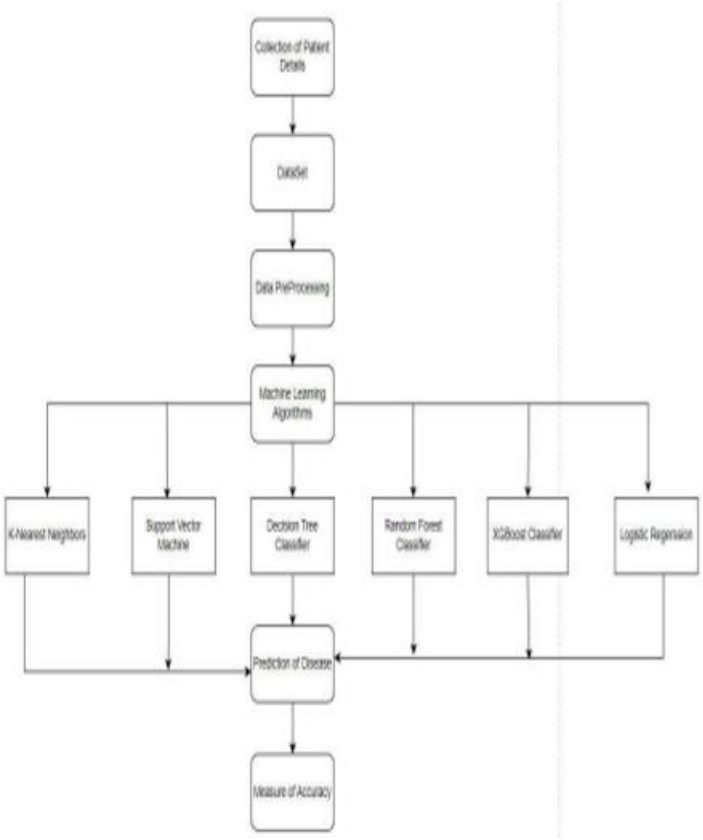


Fig 5.2

1. Raw Data Collection

Data Pre-Processing :- Before building any model, it is crucial to perform data pre-processing to feed the correct data to the model to learn and predict. Model performance depends on the quality of data feeded to the model to train. This Process includes Handling Null/Missing Values,Handling Skewed Data,Outliers Detection and Removal.

Data Cleaning:- Cleaning data involves repairing or erasing inaccurate or faulty data. a dataset's improperly structured, duplicated, or insufficient data. Remove duplicate or irrelevant observations. Filter unwanted outliers. Renaming required attributes ,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Exploratory data analysis:- Collected data utilising visual methods is called exploratory data analysis (EDA). By analytical summaries and visualisations, this is applied to identify trends, patterns, or to verify hypotheses.

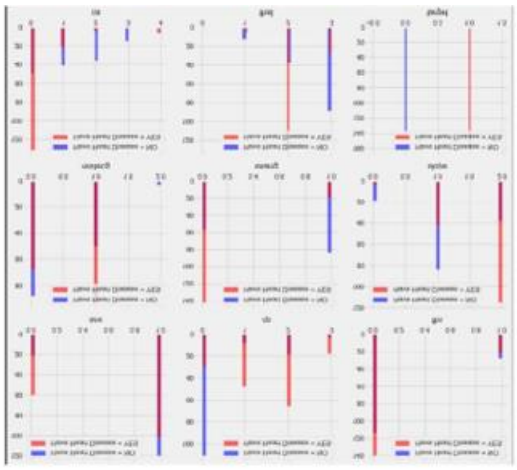
Reporting is a most important and underrated skill of a data analytics field. Because being a Data Analyst you should be good in easy and self-explanatory report because your model will be used by many stakeholders who are not from technical background.

Modelling :- Understanding metadata and their connections to other things is the method of data modelling. It is employed in the analysis of the available data needed for company processes..

System Architecture: -

The following is a description of how this system functions: Dataset collection is the act of gathering information containing patient specifics. The method of selecting attributes chooses the relevant attributes for forecasting heart disease. The available

OBSERVATIONS



The all factors as well as the objective factor exhibit significant correlations, with the exception of fbs and chol, with the weakest correlations.

Fig 6.1

Chest Pain: Those who have a cp of 1, 2, or 3 are much more inclined to develop cardiovascular disease than those who have a cp of 0.

- resting electrocardiographic results, shows that those with measure 1 (signifies an irregular heartbeat, which can range in severity from small complaints to problematic circumstances.) became likely to develop cardiovascular problem.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

(exang): Those with value 0 (No ==> exercice induced angina) had a higher risk of developing heart disease than those with value 1 (Yes ==> exercice induced angina).

- slope, or "the slope of the peak exercise ST segment": People with such a slope value of 2 (Downslopins: symptoms of an unhealthy heart) are still more likely to suffer cardiovascular disease than those with a slope value of 0 (Upsloping: improved heart rate with activity) or 1(Flatsloping: negligible change (normal).
- ca, or "number of main vessels," is a measurement of blood flow (range: 0–3). Individuals with a ca value of 0are still more prone to have cardiac disease.
- The risk of developing heart disease is higher in people with thal values of 2 (fixed defect: previously a faulty butfine now.

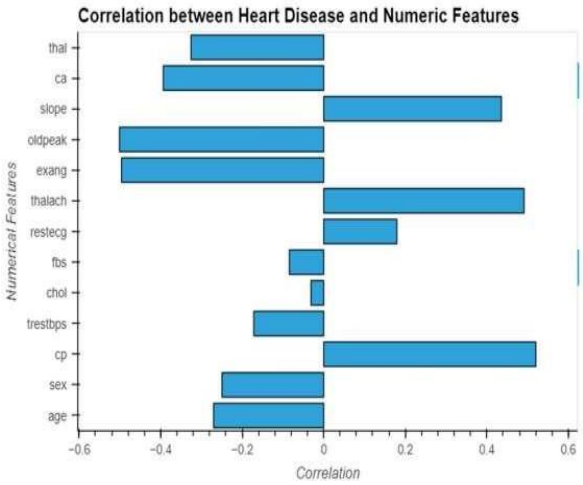


Fig 6.2

Experimental Analysis:-

We discover that the efficiency of the Logistic Regression is higher than that of other methods after executing the machine learning method for both training and testing. The confusion matrix for each method is used to calculate accuracy. Here, the number of TP, TN, FP, and FN is provided, and by applying the equation for accuracy, value has been determined. It is decided that Logistic Regression is the finest with 93.41% accuracy, and the contrast is presented below.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	92.45	93.41
1	K-nearest neighbours	88.21	91.21
2	Support Vector Machine	95.28	91.21
3	Decision Tree Classifier	100.00	83.52
4	Random Forest Classifier	100.00	91.21
5	XGBoost Classifier	100.00	91.21

Tab 1

- XGBoost Classifier, RF(Random Forest), DT(Decision Tree)are overfitting.
- Logistic regression(LR),SVM seems best models as test accuracy is more. compared to remaining models,
- KNN model performed poor compared to remaining models.

III. SUMMARY AND FUTURE IMPLEMENTATION

Deployment of important technique, Due to the fact that heart disorders are the top mortality rate in India as well as the rest of the globe, using ml algorithms for the first identification of heart issues will result in a big social influence.. Early detection of heart disease can help high-risk patients make decisions about lifestyle modifications that will lessen problems, which can be a significant improvement in thehistory of medicine. Each year, more people are diagnosed with cardiac illnesses. This calls for an early diagnosis and course of action. The medical community as well as patients may benefit greatly from the use of appropriate technology support in this area. SVM, Decision Tree, Random Forest, K- nearest Neighbors, Logistic Regression, and XGradient Boosting are some of the seven machine learning techniques utilised in this study to evaluate performance. These methods were all used in this study. The dataset, which includes 76 features, contains the expected characteristics that contribute to heart disease in individuals, and 14 significant

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

characteristics are chosen from them to help assess the method. The author receives less efficiency from the system when all the features are taken into account. Characteristic selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides greater accuracy. Certain data characteristics have virtually equal correlations, so they are eliminated. The efficiency significantly declines if every variables in the dataset are considered. The accuracy of each of the six machine learning techniques is evaluated, from which a predictions model is created. Thus, the objective is to employ a variety of evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The Logistic Regression provides the highest efficiency (93.41%) when comparing the other six.

REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9.

[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

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

(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 MutliConference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[9] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction,1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

[10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and selfreported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7. 69

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

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and geneticalgorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92. 34

[14] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

[15] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine