Industrial Engineering Journal

ISSN: 0970-2555

Volume : 53, Issue 4, No. 6, April : 2024

HEART DISEASE PREDICTION USING MACHINE LEARNING

Ms. Ankita Singh Rana, Assistant Professor, Computer Science and Engineering Quantum University Roorkee, Uttarakhand

Abstract
Machine Learning is used across many ranges around the world.The healthcare industry is no exclusion. Machine Learning can play an essential role inpredicting presence/absence of locomotors disorders, heart diseases and more. Such in formation, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing pe rpatient basis. We work on predicting possible heart diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like decision tree, Naïve Bayes, Logistic Regression, SVM and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost which can give the better accuracy and predictive analysis..

Keywords – SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix.

Overview

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm.Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

EXISTING SYSTEM

Heart disease is even being high lighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says "Prevention is better than cure", early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

- Collection, of Dataset
- Selection of attributes
- Data Pre-Processing
- Balancing of Data
- Disease Prediction

Collection of dataset

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The data set consists of 76 attributes; out of which,14 attributes are used for the system.

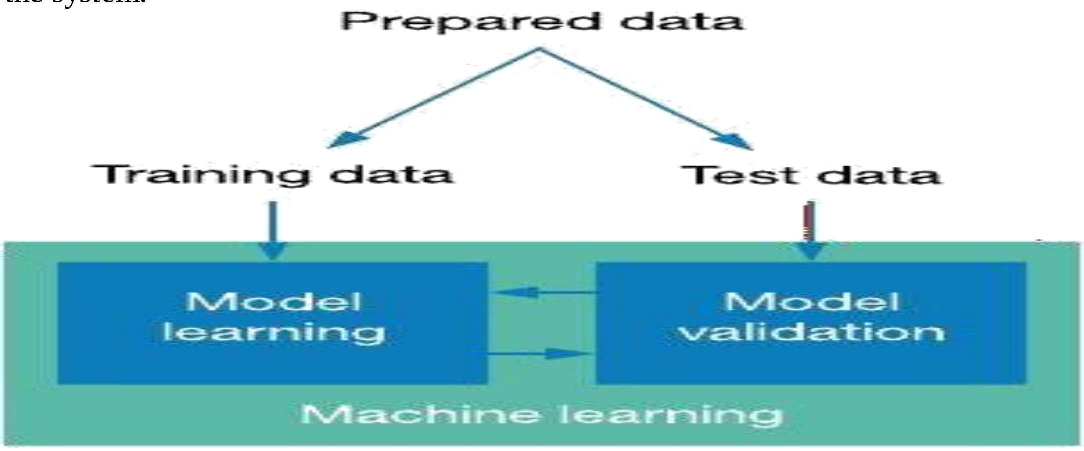
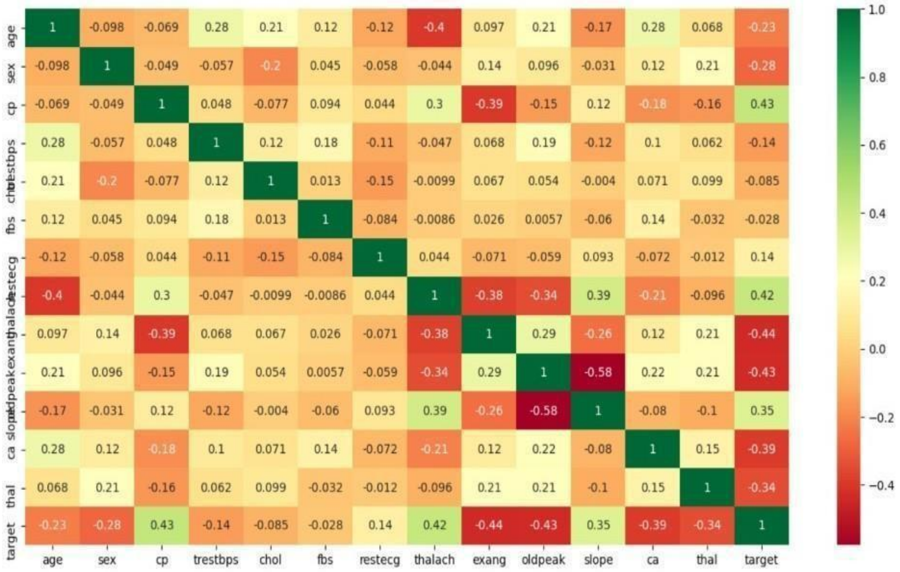


Figure: Collection of Data

Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

Figure: Correlation matrix

Pre-processing of Data

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets ,splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

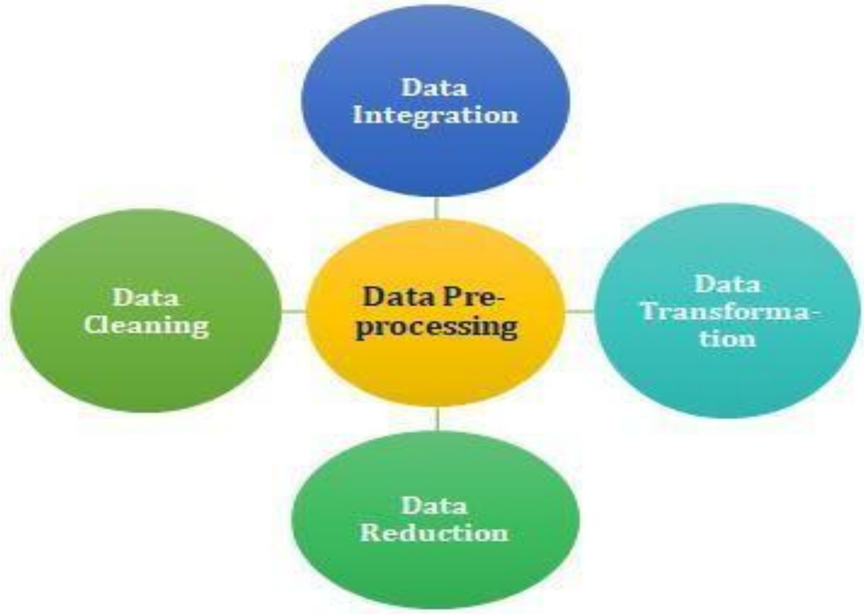


Figure: Data Pre-processing

Balancing of Data

Imbalanced dataset scan be balanced in two ways. They are Under Sampling and Over Sampling

(a) Under Sampling:

In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate.

(b) Over Sampling:

In Over Sampling, dataset balance is done by increasing the size of the scarce samples.This process is considered when the amount of data is inadequate.

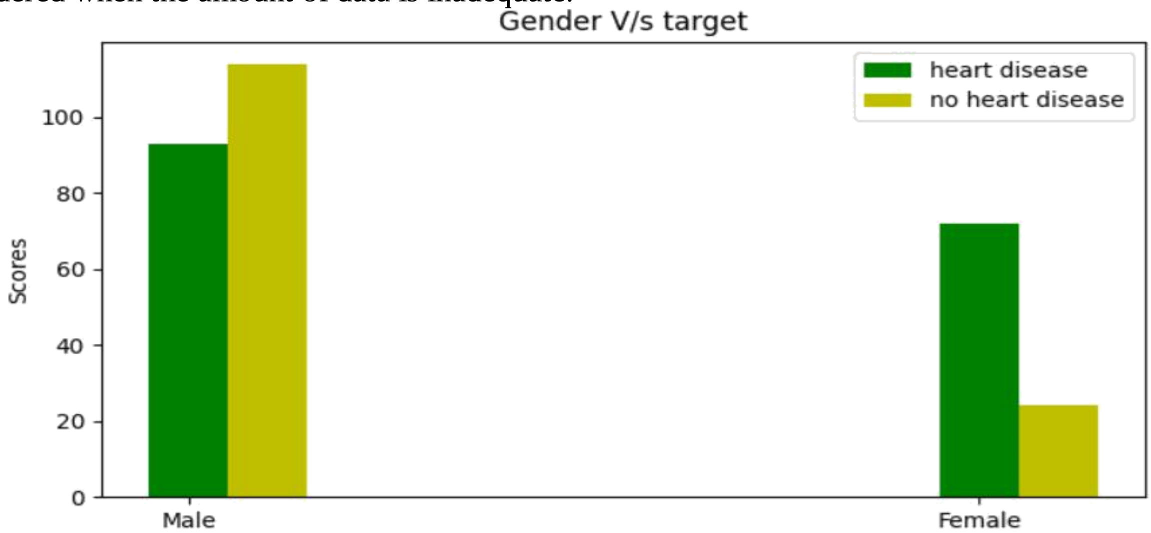


Figure: Data Balancing

Prediction of Disease

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

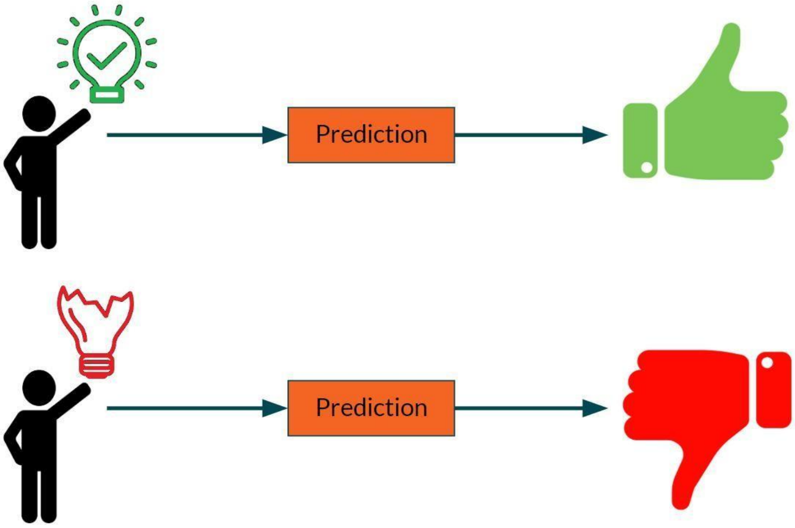


Figure: Prediction of Disease

WORKING OF SYSTEM

System Architecture

The system architecture gives an overview of the working of the system.

The working of this system is described as follows:

Data set collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on pre processed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.

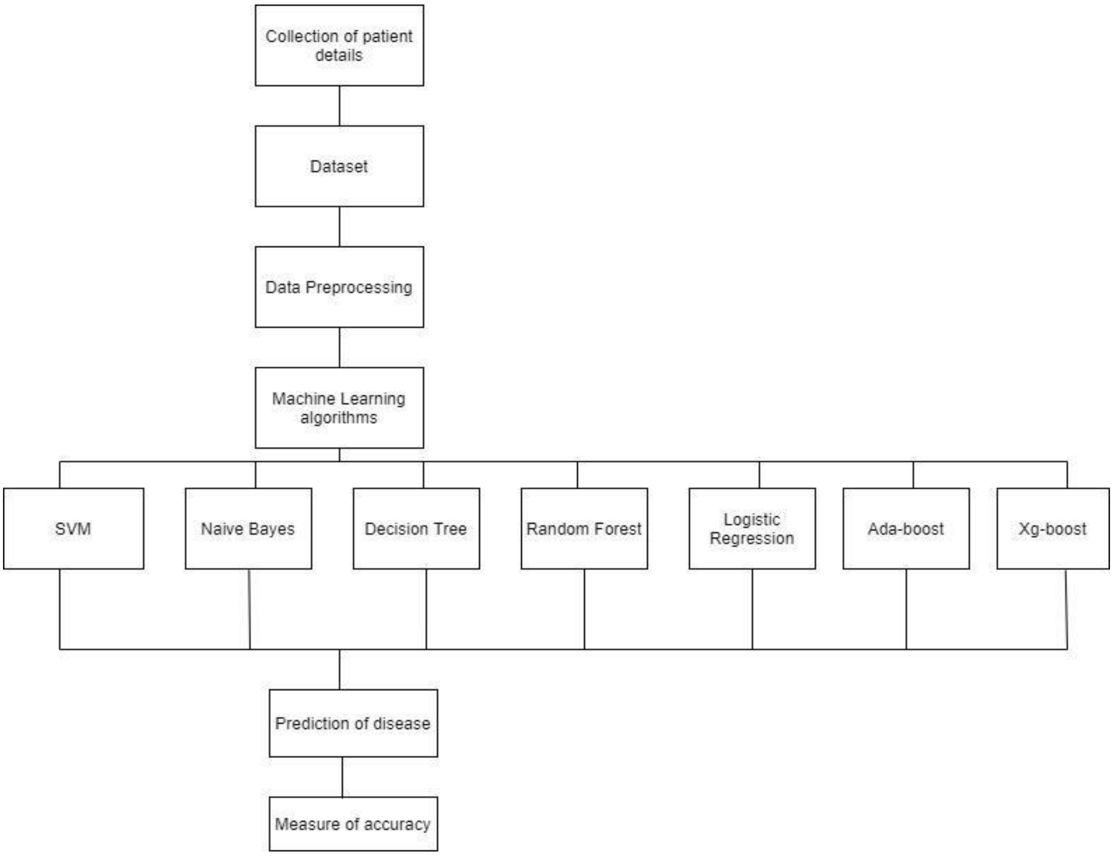


Figure:. System Architecture

Machine Learning

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

Supervised Learning

Supervised learning is the type of machine learning in which machines aretrained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

Unsupervised learning

Unsupervised learning can not be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to there al AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised

learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, weneed unsupervised learning.

Reinforcement learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training data set, it is bound to learn from its experience.

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

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.

[1] 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.

[2] 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 Scienceand Information Technologies,6(1), 637-9.

[3] 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

[4] 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

[5] 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.

[6] 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.

[7] 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 Scienceand Information Technologies,6(1), 637-9.

[8] 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

[9] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). Acoronary heart disease prediction model: the Korean Heart Study. BMJ open,4(5),e005025.

[10] 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. Arterios clerosis, thrombosis, and vascularbiology,33(9),2267-72.

[11] 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.

[12] BrownN, YoungT, GrayD, SkeneA 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.

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

[14] 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.

[15] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFI Sand genetic algorithm."International Journal of Biological, Biomedical and Medical Sciences 3.3(2008).

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

[17] 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.

[18] 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 Anti hyper tensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazo sin and chlorthali done. Current controlled trials in cardio vascular medicine