A Survey on Heart Failure prediction

[1]Moiza Shaikh, [2]Mohan Nikam

[1,2]MTech-CTIS, School of Engineering, Sandip University, Nashik, India.

[1]shaikh.moiza@gmail.com, [2]mohan.nikam@sitrc.org

***Abstract*— Due to unhealthy lifestyle practices and enormous stress in today's world , it is being completely unpredictable when a person can get a heart attack or heart failures?. Even most of the times doctors and health experts are also not able to predict the heart failures, because of this prediction of heart failure remains as a mystery even though on having so much advance technologies in the medical field. To give a try, machine learning algorithms also jump into this to predict the heart failures using some techniques like complex event processing and others. These processes involve a large amount of data to learn about the conditions and then they predict the heart failures. It is not possible to provide large amount of data every time to predict the heart failures, So some techniques should be there to provide the same in a moderate amount of datasets condition. So this paper provides a good way to predict the heart failure conditions using KNN clustering technique and Logistic regression model. Which is empowered by Artificial neural networks and Fuzzy Classification rules. Usage of Artificial neural network on clustered data, which is selective using information gain always raises the quality of the process in heart failure predictions.**

Keywords*— *Heart Failure, ANN, KNN, Fuzzy Classfication.

I. INTRODUCTION

The human heart is one of the most important parts of the body. It is responsible for providing oxygen-rich blood and vital nutrients to the different parts of the body and also performs the duty of collecting impure blood and oxygenating it in the lungs. The cardiac muscle is one of the most resilient and continuously working muscle in the body. The heart never stops, and when it does, it usually means that the person is dead. It is the heart that is responsible for the well-being of the person and is an extremely important organ in the body.

Due to this fact, any ailment or disease that occurs due to the heart can be quite critical and must be handled immediately and corrected in time. If not done on time it could prove to be very debilitating and also fatal. Therefore, the diagnosis of the heart should be done very fast and also with the highest levels of accuracy. As time is essential for the early remedial measures to be done as soon as possible and accuracy is obviously quite essential to provide the correct medication. Most of the mechanisms in place for determining heart ailments rely on manual calculations that are reliable. This is compounded by the fact that most of the scenarios of a heart disability have been quite a time critical and no time can be wasted in providing the best possible care for the patients. Therefore, a lot of research has been done to reduce the time taken to extract the results and take remedial action.

Clustering is one of the most common techniques in the realm of Machine Learning and Data Mining. Clustering refers to the technique for grouping various items according to some similar features into small groups of relevant similarity. This is very useful for application in which most of the data is contained unorganized in a large dataset. It becomes highly cumbersome to cluster this type of data manually and would also take an enormous amount of time.

K-Nearest Neighbours is one of the widely used algorithms for the purpose of clustering. It is a type of a Supervised technique, that is, the algorithm needs to be tweaked and then trained for the incoming data before its application. The K-Nearest Neighbours algorithm has a wide variety of uses; depending upon the training the algorithm can perform classification tasks or it can also perform regression tasks.

The K-Nearest Neighbours algorithm is used only when the type of data or the target variable/attribute is known before its application as it is a supervised algorithm, it needs to be trained in order to successfully cluster the data elements. The 'K' in K-Nearest Neighbours stands for the number of labels of the neighbors to be assigned to the centroids of the cluster.

Logistic Regression is one of the most common classification algorithms that is popularly used for the prediction of certain outcomes with the help of a collection of variables which aren't dependent on each other, in contrast to the method of linear regression. As Logistic Regression is a special form of Linear regression as it is categorical in its outcome variables as the odds are considered as a dependent variable. It has been used a lot since its creation in the early 20th century due to its simplicity and ease of use.

Artificial Neural Network is a network for computation that has been designed with inspiration from the workings of a human brain. Just like the Human Brain, the Artificial Neural Network has a network of tiny neurons that are layered on top of each other. These layers are interconnected with each other as the output of one layer

 DOI : 10.35291/2454-9150.2019.0064

become the input of the next layer. This is used to emulate the human brain and its numerous connections and electrical pathways made up of neurons.

This type of network is implemented in scenarios where there is a need to human-like thinking and decision-making properties. As the Artificial Neural Networks emulate the inner workings of a brain, it is highly beneficial to be implemented in critical systems such as the control systems of various critical applications such as Dams, electrical power plants, etc. This is due to the fact that they are a lot more reliable and efficient in comparison.

The term fuzzy refers to the fuzziness or the uncertainty or vagueness of some values of a certain variable. As most of the computers are run on silicon, they are by the nature of the element and the transistors that they are made up of; Binary. As we all know binary is composed of a series of 1's and 0's, or in Boolean, True or False. This is very different from the real world where there aren't completely black and white, there is gray which is somewhere between white and black.

Therefore, to represent our world a little more clearly a concept called Fuzzy is introduced, which basically accepts values that are 1 and 0 and all the values in between. This is highly useful in scenarios where absolute control is needed, such as control systems. This system has the capabilities to provide an accurate representation of the real world.

In this paper, section 2 is dedicated for the literature review of past work ,Section 3 describes the details of the developmental procedure of the model. Section 4 evaluates the results through some experiments and finally section 5 concludes this research article with the traces of the future scope.

II. LITERATURE REVIEW

A.Gavhane expresses that there has been constant growth in the incidence of heart strokes for individuals of a relatively young age. This is very problematic as experiencing a heart stroke at such a young age can be highly debilitating. Therefore, there is a need to identify the symptoms of a stroke beforehand, so that it can be prevented. As it is not feasible to do all the various tests such as blood tests and ECGs every single time. Therefore, the authors have developed a system that can predict how vulnerable a person is with regards to heart disease [1]. The system has been implemented with the help of machine learning Neural networks, with only one drawback, that is, the system is tuned only for Heart diseases and other diseases need to be added.

S. Bashar proposes a novel technique for the estimation of the heart rate of a person with the help of wearable devices. The authors state that there has been an enormous growth in the sector of Heart Rate monitoring for various purposes, such as medicine and as well as athletes. Heart Rate monitoring has been quite inconsistent with physical activity as unneeded noise and motion artifacts [2]. Therefore, the authors have presented MMMLA which utilizes Machine Learning to reduce the artifacts and achieve an accurate reading.

S. Ismaeel [3] introduces Machine Learning and its growing use in the field of medicine, especially for the prediction of heart diseases and its diagnosis. There are quite a few factors that affect the incidence of heart disease, such as age, sex blood sugar, etc. The researchers propose an ELM algorithm for the prediction of the disease and modeling of the various attributes. It is a very simple method that has the ability to do away with the expensive tests done at the hospital conventionally.

Q. Rahmanexplains the cardiovascular disease called as the Hypertrophic Cardiomyopathy. It is a debilitation disease that thickens the blood of its victim which in turn leads to poor flow of blood in the body. ECGs are widely used to diagnose this condition. The authors, therefore, develop a system capable of classifying HCM patients with the help of Machine Learning. This Machine Learning technique has been compared with the traditional means of detection and has outperformed them significantly. There is a drawback to this scheme where advanced machine learning techniques were not used. [4]

Q. Zhang elaborates that the monitoring of Heart Health is one of the most essential aspects to reduce the incidence of heart diseases. This is usually done by almost continuous tracking of the patient's ECG, but most modern devices are inconvenient and uncomfortable for the patients. Therefore, the authors designed a device that records the weak ECG from the ears with the help of a wearable device [5]. The proposed method is highly useful and considerably less intrusive as it utilizes Regression, unsupervised learning and Support Vector Machines to achieve its goal through machine learning.

M.Gjoreski[6] is concerned with the rising epidemic of chronic heart failures. As the statistics convey that the number of patients lost to cardiovascular diseases has risen consistently over the years. This is also attributed to the fact that diagnosis takes a lot of time to process. Therefore, the authors have designed a technique based on machine learning that can diagnose patients of chronic heart failure detected from the heart sounds. The Machine Learning Algorithm has been quite useful and actually produced better results than conventional methods.

A.Bhattacharya expresses that Heart auscultation is one of the most widely used and readily available methods for the detection of cardiac diseases. The method includes the use of a stethoscope and is a completely non-invasive technique. The researchers, therefore, propose a method for the detection of abnormalities of the heart with the help of a Machine Learning platform that can automatically diagnose with a wearable low-cost device. The technique has been tested to perform exceptionally in various test cases. Due to the device being low powered and portable, it makes for an exceptional device for the diagnosis. [7]

N. Omar explains the evidence-based managing tool, called a Heart Failure clinical pathway. They are very useful in the diagnosis of heart diseases, but they have one flaw, the pathway systems are static and therefore, cannot accept dynamic badges. Therefore, the authors developed a method for the dynamic clinical pathway that utilizes machine learning to provide data mining and other medical information retrieval systems [8]. The proposed system is a

 DOI : 10.35291/2454-9150.2019.0064

huge improvement over the previous version as it can now support dynamic badges and Machine Learning.

A.Batra states that in pregnant women, the main cause of concern and a cesarean section is fetal distress experienced by the fetus. This usually indicates that there is a reduction of oxygen to the fetus that could lead to a lot of complication and can severely deteriorate health while being quite fatal. Cardiotocography is the method used to diagnose this condition and also evaluate the health of the fetus [9]. The authors have utilized his technique and combined them with Machine Learning to increase the accuracy of this test to about 99.25% which is quite remarkable.

H. Bulbul [10] elaborates on the conventional technique of identifying the occurrence of Cardiac Arrhythmia in the patients with the help an ECG or the Electrocardiogram. Identifying the irregularities in the ECG graph manually is quite intensive and might also give some errors or false positives. Therefore, to automatically identify the occurrence of irregular cardiac rhythms, the authors propose a technique based on Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) which has a very high accuracy compared to the traditional techniques.

N. Dharmasiri expresses that there has been a lot of benefits to the medical fraternity with respect to technological advancements, such as Machine learning and the Internet of Things. The advancements in technology have enabled us to achieve continuous monitoring in case of patients suffering from chronic heart diseases. This is highly convenient for the majority of the patients. The authors implemented machine learning into their method and with the help of iOS devices, was able to successfully achieve monitoring. [11]

S. Radhimeenakshi introduces the concept of classification of heart diseases as it is quite essential for a medical professional to quickly identify and detect the ailment. This is usually quite time-consuming through traditional tests and diagnosis. Therefore, the authors have developed a technique [12] that utilizes Support Vector Machine and Artificial Neural Networks to achieve automated diagnosis and classification of the heart ailment. The proposed technique is highly accurate and fast in comparison with the traditional techniques.

S. Pouriyeh explains that due to the vast technological advancements in the world, there has been a lot of advancements in the medical field with the help of these innovations. The authors have implemented various Data Mining techniques in an amalgamation of Machine Learning for the prediction of Heart Diseases. The authors have evaluated the performance of various classifiers, such as Naïve Bayes, K-Nearest, Support Vector Machine and deployed them on their system [13]. The results then show that SVM is one of the best classifiers for Medical data for Heart Diseases.

L. Zhao states that there has been a lot of research done in the direction of detection of heart failure among patients. This is usually done with the help of some medical tests that determine the condition of the patient and can accurately state if the patient will have heart failure or not. As it is a matter of life and death, it is necessary to evaluate the prediction faster and with more accuracy, therefore the authors have implemented a technique for detecting heart failure with the help of Machine Learning and pulse transit time variability[14]. The proposed technique has been evaluated and produced exceptional results.

D. Kumar elaborates on the necessity of maintaining health and supporting healthcare structures to increase the quality of human life. Cardiovascular disease is one of the most fatal of the diseases as it contracts one of the most essential organs for the survival, the Heart and the accompanying blood vessels. To enable a far better evaluation of heart diseases and a proper diagnosis, the authors have implemented a machine learning framework for analyzing the massive amounts of medical data. the proposed method utilizes Naïve Bayes classifiers, Support Vector Machines, Random Forest, etc and achieves at par accuracy and performance. Due to this implementation being the first, there wasn't much fine-tuning of the parameters[15].

III. PROPOSED IDEA DESCRIPTION

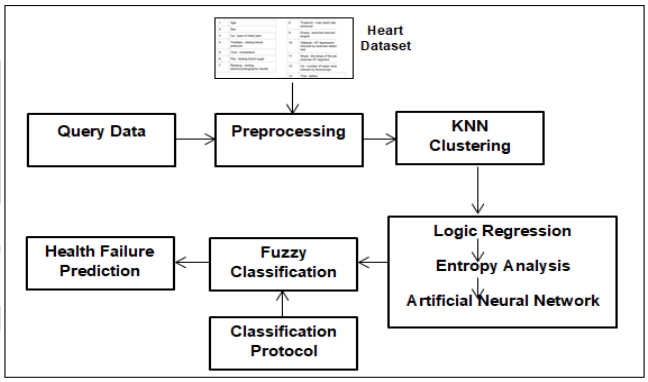

Figure 1: Overview of proposed methodology

The overview of the proposed technique is depicted in the figure1 and it is narrated using the below mentioned steps.

Step 1: Preprocessing- This is the initial step in any Machine learning model. Here initially the dataset which is obtained through the kaggle dataset forum is subjected to preprocessing.

Here in this step dataset contains many attributes like Age, Sex, Chest pain levels, Blood pressure, Cholesterol in mg/dl, Fasting blood sugar (fbs) in mg/ dl present or not (i.e is 1 or 0), History of diabetes, Maximum heart rate achieved, Exercise induced angina in 1 or 0, old peak (depression induced by exercise relative to rest), slope: the slope of the peak exercise ST segment , ca: number of major vessels (0-3) colored by flourosopy , Thal blood disorder thal: 3 = normal; 6 = fixed defect; 7 = reversible defect.

Out of all these attributes the major attributes that actually triggers the heart failure or heart attack conditions are selected to form a list. The major attributes which are selected are Chest pain levels, Blood pressure, Cholesterol in mg/dl and Maximum heart rate achieved.

Step 2: K-nearest Neighbor clustering- KNN or K-nearest neighbor clustering technique is applied to cluster the training dataset for the preprocessed four attributes as mentioned in the earlier step. Initially, each and every row of the dataset is measured for the Euclidean distance with

 DOI : 10.35291/2454-9150.2019.0064

respect to all other rows and then mean of this distance is assigned as the row Euclidean distance R_{ED}. Then the average of this row Euclidean distance R_{ED} is measured to yield the Euclidean distance E_D of the complete preprocessed dataset.

All the rows which are appended with their respective distances are sorted in ascending order to bring the nearest neighbors more closer. Then, based on the random data points cluster boundaries are being evaluated with the help of E_D. Based on these boundaries, clusters are formed and labeled in well maintained list.

Step 3: Logistic Regression and Entropy Analysis - The each and every cluster from the past step is subjected to estimate the feature and the probability of the particular outcomes. As we know that the outcome data is not yet been categorized, So here in this step a abstract probability of the data is estimated using the mean and standard deviation of the R_{ED} through equations 1,2 and 3 for all the clusters.

$$\mu = \frac{(\sum_{i=1}^{n} xi)}{n} \quad \text{_____(1)}$$

$$\delta = \sqrt{\frac{1}{N}} \sum_{i=1}^{n} (xi - \mu)^2 \quad \text{_____(2)}$$

$$Pd = \sum_{i=0}^{n} xi \sum_{\mu+\delta}^{n} REd \quad \text{_____(3)}$$

Where ,

μ - Mean of the Row Euclidean distance R_{ED}

δ- Standard deviation

Pd - Probability cluster data

Once this step of the regression is over, then the newly formed regression clusters are used to estimate the distribution factors of heart failure probability for the given test data using the information gain theory. Where each and every revised cluster are counting for the probability of heart disease with respect to the maximum values of the heart failure proneness. And by using the information gain theory clusters are selected for the non zero values of the gain. The Information gain can be estimated using the equation 4.

$$IG = -\frac{P}{T}\log\frac{P}{T} - \frac{N}{T}\log\frac{N}{T} \quad \text{_____(4)}$$

Where

P= Frequency of the probability count

N= Non probability count

T= Cluster Elements Size.

IG = Information Gain of the cluster

Step 4: ANN and Fuzzy Classification - The selected clusters from the information gain theory is used to form neurons here based on the mean and standard deviation of the R_{ED}. Every cluster is converted into 3 neurons. First neuron is formed for the range : $RED < (\mu - \delta)$, Second neuron is formed for the range $RED >= (\mu - \delta)\ AND\ RED <= (\mu + \delta)$ and finally third neuron is formed for the range : $RED > (\mu - \delta)$.

Each of the formed neuron is now dealing with the heart attack or failure classification protocols to label the respective rows. And then these rows are being counted and forward to Fuzzy Classification theory to classify the outcome for the proneness of the Heart failure or not.

In the fuzzy classification theory first the maximum count is being stored in the database to learn the future outcomes. Every time this maximum value is fetched from the database from the past learned history. And then it is divided in 5 crisp values set like VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH. The present count of the test data is being evaluated in all the 5 ranges to detect the probability of the heart failure or heart attack proneness.

IV. RESULT AND DISCUSSIONS

The proposed method of heart attack or heart failure detection is deployed in windows machine with the configuration of Corei5 Processor and 6GB of primary memory. The model is developed in the Java programming language by using Netbeans as Integrated Development Environment along with the Mysql database . For the deployment of the model proposed system uses the heart failure Dataset from the open dataset forum like Kaggle and its attributes are described in the past sections. Some experiments are carried out to measure the effectiveness of the proposed model as stated below.

Time Comparison : When the Time comparison of the KNN of the proposed model is compared with that of Naive Bayes clustering algorithm of [16] we got some results as shown in the table 1. Then on observing the graph in figure 2 plotted for the values of table 1. It clearly indicates that the KNN of the proposed model takes considerable less time than that of naive Bayes clustering. The results shows that Naive bayes algorithm takes around average of 65 Milliseconds where as KNN takes average of around 47.75 seconds.

Experiment No	Training Dataset	Testing Dataset	Naïve bayes (Time in milli seconds)	KNN (Time in milli seconds)
1	80	20	30	25
2	70	30	90	44
3	60	40	20	57
4	50	50	120	65

Table 1: Time Required for the execution

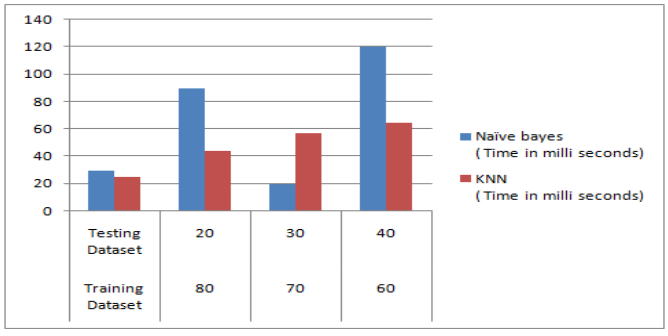

Figure 2:Time comparison for time of execution

Precision and Recall - Precision and recall are the most effective measuring parameters in prediction mechanisms. They can be more effectively stated as below.

A = The number of correctly predicted values

B= The number of incorrectly predicted values

 DOI : 10.35291/2454-9150.2019.0064

C = The number of correctly predicted values are not predicted

So precision can be given as

Precision = (A / (A+ B)) *100

Recall = (A / (A+ C)) *100

When the precision and recall values are recorded for some set of experiments as shown in table 2. we found that the average values of precision and recall are around 70%, Which is higher than that of [16]. [16] uses the 10 input factors with 500 iterations using CNN. Whereas the proposed model uses the 14 input factors and it is deployed using the ANN along with the fuzzy classification theory. And the figure 3 clearly indicates that proposed model which is using ANN and Fuzzy technique over performs than that of CNN_UDRP model of [16].

Table 2: Precision and Recall

No of Testing data	Relevant predictions identified (A)	Irrelevant Predictions Identified (B)	Relevant Prediction not identified (C)	Precision in %	Recall in %
25	17	8	8	68	68
50	32	18	18	64	64
75	55	20	20	73.33333333	73.33333333
100	77	23	23	77	77

Table 3: Performance of CNN_UDRP and ANN Fuzzy

Parameters	ANN- Fuzzy	CNN _ UDRP
Precision	70.58	62
Recall	70.58	60
Accuracy	70.58	65
F Measure	72.4	59

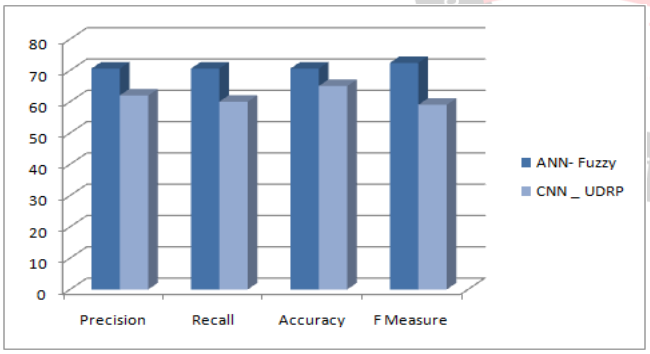

Figure 3: Performance of CNN_UDRP and ANN Fuzzy

V. CONCLUSION AND FUTURESCOPE

The proposed model of heart failure prediction extensively uses the KNN clustering algorithm . And this is powered with the fuzzy classification model and ANN. The model follows all the novel rules to dugout the proper prediction for the given input testing data. On performing several experiments and evaluation technique, it is found that the proposed model performs well to produce around 70% of the accuracy that is really good in the very first attempt of the deploying the model.

In the future this model can be enhanced to work on more attributes collected from any hospital or research labs to help real time patients though a mobile application.

REFERENCES

[1] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning", Proceedings of the 2nd International Conference on Electronics, Communication, and Aerospace Technology (ICECA 2018).

[2] Shikder Shafiul Bashar et al, "A Machine Learning Approach for Heart Rate Estimation from PPG Signal using Random Forest Regression Algorithm", International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February 2019.

[3] Salam Ismaeel, Ali Miri and Dharmendra Chourishi, "Using the Extreme Learning Machine (ELM)Technique for Heart Disease Diagnosis", IEEE Canada International Humanitarian Technology Conference (IHTC), 2015.

[4] QuaziAbidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and HagitShatkay, "Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", IEEE Transactions on Nano Bioscience, 2015.

[5] Qingxue Zhang, Dian Zhou, and Xuan Zeng, "Hear the Heart: Daily Cardiac Health Monitoring Using Ear-ECG and Machine Learning", IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017.

[6] Martin Gjoreski et al, "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers", 13th International Conference on Intelligent Environments, 2017.

[7] Anitek Bhattacharya, Mohan Mishra, Anushikha Singh & Malay Kishore Dutta, "Machine Learning Based Portable Device for Detection of Cardiac Abnormality", International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT), 2017.

[8] N. Omar et al, "Personalized Clinical Pathway for Heart Failure Management", International Conference on Applied Engineering (ICAE), 2018.

[9] A. Batra, A. Chandra and V. Matoria, "Cardiotocography Analysis Using Conjunction of Machine Learning Algorithms", International Conference on Machine Vision and Information Technology, 2017.

[10] H. Bulbul, N. Usta and M. Yildiz, "Classification of ECG Arrhythmia with Machine Learning Techniques", IEEE International Conference on Machine Learning and Applications, 2017.

[11] N.D.K.G. Dharmasiri and S. Vasanthapriyan, "Approach to Heart Diseases Diagnosis and Monitoring through Machine Learning and iOS Mobile Application", International Conference on Advances in ICT for Emerging Regions (ICTer), 2018.

[12] S. Radhimeenakshi, "Classification and prediction of Heart Disease Risk using Data Mining of Support Vector Machine and Artificial Neural Network", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.

[13] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sanninoy, Giuseppe De Pietroy, Hamid Arabnia, and Juan Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", IEEE Symposium on Computers and Communication, 2017.

[14] Lina Zhao, Chengyu Liu, Shoushui Wei, Changchun Liu, and Jianqing Li, "Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning", IEEE Access (Volume: 7), 2019.

[15] Dinesh Kumar G, Santhosh Kumar D, Arumugaraj K, and Mareeswari V, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, 2018.

[16] Sayali Ambekar and Rashmi Phalnikar, " Disease Risk Prediction by Using Convolutional Neural Network" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),IEEE,2018.

 DOI : 10.35291/2454-9150.2019.0064