Heart of Prediction by using Hidden Markov model and Fuzzy Classification

Jaybhaye Nagin Madhukar[1] Ghorpade Vrushali Shamrao[2] Katake Rutuja Maruti[3] Kathile Shubhangi Jagannath[4] Prof. M.P.Kharche[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]KJCOEMR , Pune, India

Abstract— In today's fast-paced world, a lot of individuals are not concentrating on their health as much and have increasing amounts of stress which lead to the introduction of a large number of unhealthy lifestyle choices that have been affecting their health significantly. These factors contribute to the large-scale increase in the number of heart diseases and failures which are highly difficult to predict. Even the doctors are unable to perform any predictions as they are unable to process the varied parameters that govern the rate of heart failures. Therefore, the machine learning paradigm comes to the rescue in this situation as it can analyze and process complex events to gain predictions. The machine learning approaches require massive amounts of data to enable complex processing, which is not ideally possible due to the inherent nature of personal medical data, due to which a sparse number of datasets are available in this regard. Therefore, this publication defines an innovative solution that utilizes Fuzzy C-means clustering, Logistic Regression along with Hidden Markov Model and Fuzzy Logic to achieve effective and accurate Heart Failure predictions. The proposed methodology has been rigorously tested against conventional approaches that have produced exceptional results and prove the increased accuracy and speed of the technique defined in this paper.

Keywords: Fuzzy C Means Clustering, Hidden Morkov Models, Fuzzy Classification, Hear Disease Prediction

I. INTRODUCTION

Health what are the most important aspects of human life. Health is important so that human beings can perform at their peak productivity at all times. This is only possible when human beings have good health and various techniques that can maintain this state for a long time. Health can be maintained through various advances that have been done in the medical field. There have been critical advancements in medical care throughout human history. Years of research and hard work has led to great advancements in maintaining good health. Various contribution throughout the centuries has allowed us to reach the epitome of peak human health.

These advancements in health have a loud for a greater quality of life among humans. Better health care and Healthcare systems have allowed for a much better lifestyle which is free from pain and suffering. Unlike the early and middle ages which were filled with various diseases and ailments that ravaged human beings. Several diseases have been attributed to large-scale deaths that have been due to the inability of the medical system at that time to understand the disease and provide for a cure. Humans have come a long way from those days and the Healthcare system has evolved significantly. This has led to a marked decrease in infant mortality as well as a significant increase in the average life expectancy.

This has led to a marked decrease in the attention that people pay towards their health. Coupled with the increasing aging population which leads to a very deadly combination that is difficult to be dealt with conventional approaches. There have also been significant Lifestyle changes and practices that have been gaining popularity in the masses. These lifestyle changes are highly unhealthy and could lead to disastrous results in the form of heart failures or heart attacks. Due to changes in lifestyle, it becomes highly difficult to protect when a person can get Heart Attack or heart failure. This is the reason the prediction of heart attacks and failures has been highly difficult and a complex endeavor even though there have been significant advancements made in the medical field.

For the purpose of predicting heart attacks of heart failure machine learning algorithms are the best choice. This is due to the fact that machine learning algorithms can utilize various techniques and parameters to solve this complex event. Machine learning algorithms are highly useful as they can process a large amount of data and extract insightful information from that data easily. For the purpose of prediction where is machine learning algorithms require past data for the purpose of training of the algorithm. After the training face, the algorithm can be applied to actual real data to provide accurate and efficient predictions. Most of the time this training data which is required in large amounts is not easily available. This leads to a decrease in the accuracy of the algorithm in predicting heart failure.

Therefore, the machine learning paradigm must be modified in a way to accept small amounts of data and utilize it for the purpose of prediction. The accuracy of the algorithm needs to be significantly high as having fewer data would lead to you lot of false positives and negatives to creep into the results. The machine learning algorithm that is the best suited for predicting heart attacks and heart failures is the hidden Markov model or HMM.

Hidden Markov models are statistical models that are used for enhanced predictions through a Markov process. The hidden Markov models can be understood by comparing them representing them as a Bayesian network that is dynamic and simple in nature. This allows for the modeling of the Markov model that is hidden or unobservable. The Markov models are time series models that allow for classification in terms of the various parameters that govern the possibility of a heart attack or heart failure. The hidden Markov models can allow for determination of the hidden state which is the possibility of a heart attack or heart failure accurately. This machine learning approach is catalyzed with the use of Fuzzy Logic.

The introduction of Fuzzy C-Means along with Logistic regression allows for the evaluation of Fuzzy crisp

values that can be utilized for providing classification labels. The Logistic regression processes the small amount of data that is given as an input. The process data is analyzed by the hidden Markov model and then the output is given to the Fuzzy Logic module for the creation of is then rules which lead to the ultimate classification labeling. These labels are the indicators of the predictors of the event of heart attack or heart failure. Therefore the introduction of a machine learning algorithm can facilitate highly accurate predictions of heart attack and heart failure.

In this research article section 2 is dedicated to Literature Survey whereas Section 3 is designed to elaborate the Proposed methodology ideas. Section 4 discusses the results and finally the section 5 concludes this paper along with the future work trails.

II. LITERATURE SURVEY

A. Daoudy [1] expresses from the last few years the healthcare analytics has become an important issue for a researcher. There is various experiment implemented for the detection of heart attacks and other various diseases. Use healthcare data on the cloud for clustering is used for map-reduce using K-means clustering on the cloud. Over big data from healthcare communities convolutional neural network machine learning is used disease prediction. In the proposed paper implements the heart disease monitoring system using Spark and Cassandra frameworks. This framework observes a real-time classification model on heart disease attributes of patients. Thus, the results of the proposed paper are a simpler and more effective way.

G. D'Addio [2] narrates from the last decade there is a huge increase in the number of patients due to heart disease. Heart rate fluctuation is known as heart rate turbulence (HRT) can be expressed as turbulence onset (TO) and slope (TS). The physiologic pattern of HRT contains of a momentary heart rate acceleration attend by a more gradual heart rate deceleration before the rate returns to a pre-ectopic level. The proposed paper advises not to use mean instead of that they can use the median value of HRT indexes because of the reliability of the turbulence slope that appears greater than the turbulence onset index.

A. Nainwal [3] explains one of the major causes of an unhealthy state of mind all over the world is Congestive heart failure (CHF). Despite improvements in medical therapies and surgical techniques, there is larger growth in heat failure patients. To increase in both prevalence and incidence pump failure is a common and major health problem worldwide. The pumping process of CHF affected heart is moderate than a normal heart. In the proposed paper, the analysis shows out of ten features six features have strongly significant changes. Which is considered for a doctor to determine the heart failure from the patient's ECG.

L. Pecchia [4] specifies early detection of CHF is asymptomatic is crucial to avoid the condition worsening and to prevent the complication of clinical conditions, which may cause higher social costs. The main aim of this paper is to aim is by applying classification and regression tree (CART) to look over the power of short-term HRV features in classifying CHF patients according to disease severity. The whole process of classification is as follows 1) Excerpts Classification 2) Subject Classification 3) Validation. The excerpts are constituted by RMSSD, TOTPWR, HF, and LF/HF achieves the best results in classifying.

R. Maestri [5] proposes the existence and the nature of rhythmic changes in EEG related to ventilator oscillations in heart failure (HF) patients with periodic breathing (PB). Periodic breathing (PB) is a cyclic rise and fall in ventilation in which hyperventilation phases are separated by periods of apnea or hypopnea. Both during daytime and nighttime (prevalence: 50-85%) breathing disorder very common in heart failure (HF) patients. Due to a reduction in hemodynamics which determines slow circulation time between lungs occurs in patients, this loss of stability is very probable.

C. Chen [6] narrates that all over the world heart failure (HF) is an increasing health problem. In the United States, more than 5 million people suffered from HF and for HF patients each year 30 billion dollars are required on medical treatment. Thus, in the proposed paper the author performs statistical analysis and verified prediction results by using the clustered group from CGMH patient medical record by using some of the features mentioned in the record. Heart Failure patients can be divided into four major subgroups including systolic, diastolic, valvular, and non-specified types. The accuracy rate of the proposed paper is predicted by 75.26% through a lOfolds cross-validation mechanism.

R. Rosli [7] aims that hospitals use a continuous cardiac monitoring machine. Immobile, heavy, costly, big and less alarming to monitor the changes in heart rate of a patient with potentially fatal heart diseases. There are several symptoms of heart diseases such as Chest, arm or below breastbone pain and discomfort, Sweating, vomiting, nausea or dizziness, Fatigue or shortness of breath, Rapid or irregular heartbeats. The purpose of the paper has been successfully implemented. 53% of them say that it is very appropriate to be one of the serving factors for early warning of Potentially-Fatal Heart Diseases (such as heart attacks, heart failures, etc.).

L. Baumann [8] explains to enhance cardiac resynchronization therapy (CRT) for chronic heart failure (CHF) patients the pulse pressure increases of left ventricular pacing site and atria-ventricular (AV). The main of this paper is of ventricular conduction delay. the major steps taken by the author are data collection, pacing treatments, ACL optimization algorithm, ACL algorithm validation, QRS width criterion. Two distinct clinical advantages of the ACL optimization algorithm are it provides permission to physicians to observe pacing therapy for PP increase without using an invasive PP sensor and second is the ACL algorithm allows physicians to check many combinations of pacing sites and AV delay quickly by utilizing transient rather than steady-state measurements.

V. Rolle [9] analysis heart failure (HF) is a multifactorial pathology presenting one of the highest prevalence worldwide. There are different therapies are provided to patients suffering from HF, including changes in lifestyle, pharmacological agents' vasodilators, diuretics, beta-blockers. For the substitution process, it requires properties of the original model to preserve the numerical

and physiological in coupling transformations second is parameter identification of coupling transformations and third is the impact of integrating the new pulsatile model by providing information. Thus, the CRT devices will be useful to tackle this issue.

K. Weigand [10] introduces ventricular tachycardia (VT) and ventricular fibrillation (VF) are the most familiar causes of unexpected cardiac death in patients with chronic heart failure (CHF). In pre-clinical translational research absorption and executing EP testing in this model may be valuable. CHF rat model using modified clinical catheters and electrodes is introduced by the proposed paper. Currently, there is no effective clinical technique to control if a drug has the potential to induce VT in an animal model before it is tested on patients. The electrophysiologic analysis could be used to determine the arrhythmogenic potential of new treatments for CHF

N. Arzeno [11] expresses over the past decade as a mortality predictor the congestive heart failure (CHF) population has become an interesting topic to be researched. To predict mortality in congestive heart failure (CHF) linear and nonlinear indices of heart rate variability (HRV) have been shown in the proposed paper. In nonlinear indices, it only describes the fractality or complexity of HRV. In linear time- and frequency-domain), complexity (sample entropy), fractal (detrended fluctuation analysis) and chaos (numerical titration) analyses on the HRV are performed over the database fifty patients.

G. Ferrari [12] elaborates decreased cardiac function produces an imbalance between metabolic demand of peripheral tissues and cardiac output in Congestive Heart Failure (CHF) both acute myocardial ischemia and in chronic heart failure are protective during Vagal stimulation (VS). The surgical plan of action is uneventful in all patients. The beneficial effect is analyzed in patients with heart failure by several mechanisms. Chronic vagal stimulation is a further treatment option for CHF patients and it should be integrated with defibrillator and resynchronization therapies.

S. Babu [13] narrates the modern lifestyle health diseases are increasing tremendously. Our health causing heart diseases and other health problems it depends on our lifestyle. Thus, the process of discovering actionable information from a large set of data in data mining is an advanced technology that analyzes large volumes of data and extracts patterns for important outcomes. The data is collected contains the fourteen attributes that are extracted such as age, sex, blood pressure, and blood sugar, etc. can predict the likelihood of a patient getting heart disease. Thus this paper proposed early detection of heart disease and its diagnosis correctly on time.

III. PROPOSED IDEA DESCRIPTION

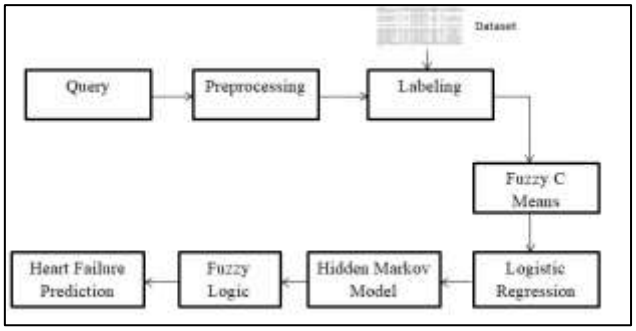

Fig. 1: Overview of the proposed methodology

The proposed methodology has been elaborated in detail in the system overview image given in figure 1 above. The procedure of the presented technique is elaborated further in the steps given below.

1) Step 1: Preprocessing –

The preprocessing is the first and foremost step in this procedure as it is a highly necessary feature to condition the dataset acquired from the Kaggle portal to facilitate machine learning implementations.

The dataset contains a combination of attributes such as Thal blood disorder of 3 types thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, ca: number of major vessels (0-3) that are colored by fluoroscopy, slope: which is the slope of the peak exercise ST segment, old peak, Exercise-induced angina present or not, Maximum heart rate recorded, any present history of diabetes, fasting blood sugar (fbs) and Cholesterol in mg/ dl, Blood pressure, Chest pain levels, along with the attributes of the patient such as age, sex, etc.

These are a lot of attributes and from these attributes, the most significant attributes that can cause heart failure are selected to create a list. The selected attributes range from, Maximum heart rate achieved, Cholesterol in mg/dl, Blood pressure to Chest pain levels.

2) Step 2: Fuzzy C-means clustering –

In this step the preprocessed list containing the selected attributes is taken as an input for the cluster formation through the use of the Fuzzy C-means clustering technique. First and foremost, the Euclidean Distance of each of the rows is calculated in correspondence to the other rows. This distance is then utilized to achieve the mean of all the rows combined and is stored as the row Euclidean distance R_{ED}. The Euclidean Distance of the complete dataset is then calculated by taking the average of all the row Euclidean distance R_{ED}.

The rows are then sorted in ascending order by utilizing the Euclidean distance E_D that is appended to the respective rows. For the formation of the clusters, some boundary conditions need to be assigned therefore Fuzzy Crisp value data points are selected. These data points serve an important role in the assignment of the clusters and their labels which are subsequently stored in a list.

3) Step 3: Logistic Regression and Entropy Analysis –

The clusters obtained in the previous step are given as an input to perform logistic regression. The logistic regression achieves the purpose of estimating the probability and the features of the data. This is done through the utilization of the mean and standard deviation on the row Euclidean

distance R_{ED} through the equations given in 1 and 2 which estimates an abstract probability of the data.

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

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

Where,
AVG - Mean of the Row Euclidean distance R_{ED}
δ- Standard deviation
Pd - Probability cluster data

Once the regression list is created using the logistic regression module, the regression clusters are utilized to achieve the distribution of the factors that decide the probability of heart failure in the subject. This is achieved through the use of entropy estimation by implementing the Shannon information gain given in equation 3 below for the logarithmic regression ratios. The gain values are then selected for each of the clusters with a non-zero result with maximum proneness to heart failure and added to the information gain feature list.

$$IG = -\frac{X}{Z}\log\frac{X}{Z} - \frac{Y}{Z}\log\frac{Y}{Z} \quad \text{_____} \quad (3)$$

Where
X= Frequency of the probability count
Y= Non-probability count
Z= Cluster Elements Size.
IG = Information Gain of the cluster

ALGORITHM 1: Regression Analysis through Logarithmic Ratios

```
//Input : FCM Clusters FC_L , Prediction Factor P_F
//Output: Entropy List Of clusters ENT_L
1: Start
2:     FOR i=0 to Size of FC_L
3:     SG_L = FC_L [i] [SG_L = Single Cluster]
4:     T_LST = Ø [T_LST = Temp List]
5:      count=0
6:          FOR j=0 to Size of SG_L
7:          R_L = SG_L [j] [R_L = Row List]
8:          SUM=0
9:              FOR k=0 to Size of R_L
10:             SUM=SUM+R_L[k]
11:             END FOR
12:          IF( SUM > P_F )
13:          count++
14:          End FOR
15:     X=count, Z= Size of SG_L, Y = Z-Y
16:     E= (-X/Z) log(X/Z) (-Y/Z) log(Y/Z)
17:     I_LST[0]=I , I_LST[1]=E
18:     ENT_L = ENT_L + I_LST
19:     END FOR
20: return ENT_L
21: Stop
```

4) Step 4: Hidden Markov Model and Fuzzy Classification

The information gain features obtained in the previous step are utilized in this step for the actual prediction of heart failure. The mean and standard deviation of the row Euclidean distance R_{ED} are calculated along with the information gain to create neurons. Each of the clusters is utilized to form 3 neurons that are based on the maximum and minimum ranges obtained through the comparison between the information gain values and the mean and standard deviation values of the row Euclidean distance R_{ED}.

The maximum count is stored in the database for the fuzzy classification purpose for future reference in the subsequent outcomes. As a new value is encountered, it is compared with the maximum count and classified using 5 crisp value set such as VERY HIGH, HIGH, MEDIUM, LOW, VERY LOW.

This gives a wide range of values that are effectively classified to achieve the highly accurate Heart failure proneness prediction.

IV. RESULT AND DISCUSSIONS

The presented technique for heart failure prediction has been developed on a machine with a standard configuration. The development machine consisted of an Intel Core i5 processor coupled with 4GB of RAM and 500GB of storage. The proposed methodology has been coded in the Java programming language utilizing the NetBeans 8.0 IDE and the Database requirements are fulfilled by the MySQL database.

The dataset utilized for this methodology has been obtained from the Kaggle open dataset forum. For determining the performance metrics of the presented technique, a collection of experiments is performed which are outline below.

Precision and Recall - The precision and recall performance metrics are one of the most accurate markers of performance that can enable effective evaluation of the proposed methodology. The parameters are elaborated mathematically in the equations given below.
A = The number of correctly predicted values
B = The number of incorrectly predicted values
C = The number of correctly predicted values are not predicted
D = The number of Correctly Predicted Incorrect values
So, precision can be given as
Precision = (A / (A+ B)) *100
Recall = (A / (A+ C)) *100
Accuracy = (A+D)/ (A+B+C)
F-Measure = (2*Precision*Recall)/ (Precision + Recall)

The precision and recall metrics are measured in the utmost detail for the proposed methodology and tabulated in table 1 given below. We notice that the average precision and recall achieved for the presented methodology is around 70% which is significantly higher than the values achieved in [14]. The approach in [14] deploys the CNN with 500 iterations and 10 input values. The presented technique, on the other hand, employs the Hidden Markov Model and Fuzzy Logic along with 14 input factors.

The comparison of the methodologies is tabulated in Table 2 and the values are subsequently plotted in the bar graph given in figure 2 below. The graph indicates that the presented technique for Heart Failure prediction utilizing the Hidden Markov Model and Fuzzy Logic outperforms the CNN_UNDRP model utilized in [14].

No of Testing Data	Relevant Predictions Identified (A)	Irrelevant Predictions Identified (B)	Relevant Predictions Not Identified (C)	Correctly Predicted Incorrect values (D)	Precision	Recall	Accuracy
25	11	4	5	5	73.33333333	68.75	80
50	24	8	12	6	75	66.66666667	68.18182
75	41	15	13	6	73.21428571	75.92592593	68.11594
100	59	21	16	5	73.75	78.66666667	66.66667
125	71	26	23	5	73.19587629	75.53191489	63.33333

Table 1: Precision, Recall & Accuracy

Parameters	HMM-Fuzzy	CNN_UDRP
Precision	76.85	62
Recall	76.49	60
Accuracy	69.25	65
F Measure	76.66	59

Table 2: Performance of HMM Fuzzy and CNN_UDRP

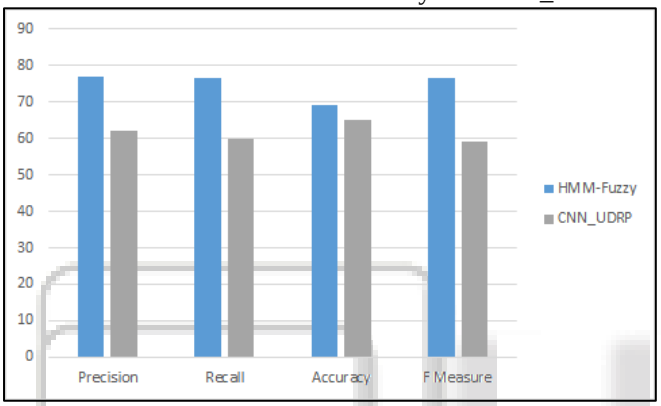

Fig. 2: Performance of HMM Fuzzy and CNN_UDRP

V. CONCLUSION AND FUTURE SCOPE

The presented technique for the prediction of heart failure in individuals utilizes the Fuzzy C-Means algorithm extensively for clustering. The prediction model is assisted by the Hidden Markov Model and the Fuzzy logic approach. The testing of the proposed methodology yields significantly positive results that outperforms the traditional approaches in terms of speed and accuracy by a large margin. The extensive experimentation reveals that the accuracy achieved in the proposed methodology is a good indication in the first execution attempt of deployment.

For future research purposes, a mobile application can be developed for a real-time scenario which can help the doctor screen patients readily and the number of attributes can be increased to further increase the accuracy.

REFERENCES

[1] Abderrahmane Ed-daoudy, Khalil Maalmi," Real-time machine learning for early detection of heart disease using big data approach" 978-1-5386-7850-3/19/, IEEE 2019.

[2] Gianni D'Addio, Mario Cesarelli, Graziamaria Corbi, Maria Romano, Giuseppe Furgi, Nicola Ferrara, Franco Rengo," Reproducibility of Heart Rate Turbulence Indexes in Heart Failure Patients" 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, August 31 - September 4, 2010.

[3] Ashish Nainwal, Yatindra Kumar, Bhola Jha," Morphological Changes in Congestive Heart Failure ECG" 978-1-5090-3480-2/16/.00 ,IEEE 2016.

[4] Leandro Pecchia, Paolo Melillo, Mario Sansone, and Marcello Bracale," Discrimination Power of Short-Term Heart Rate Variability Measures for CHF Assessment" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 15, NO. 1, JANUARY 2011.

[5] Roberto Maestri, Maria Teresa La Rovere, Elena Robbi, Gian Domenico Pinna," Relationship Between Ventilatory Oscillations and Fractal Dimension of the EEG During Daytime Periodic Breathing in Heart Failure Patients" 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.

[6] Chi-Jim Chen, Ying-Tsang Lo, Jhen-Li Huang, Tun-Wen Pail ', Min-Hui Liu, Chao-Hung Wang," Feature Analysis on Heart Failure Classes and Associated Medications" IEEE International Conference on Systems, Man, and Cybernetics' SMC 20161 October 9-12, 2016' Budapest, Hungary 2016.

[7] Rafhanah Shazwani Binti Rosli, Rashidah Funke Olanrewaju," Mobile Heart Rate Detection System (MoHeRDS) for Early Warning of Potentially- Fatal Heart Diseases" International Conference on Computer & Communication Engineering 2016.

[8] Lawrence Baumann, Veerichetty Kadhiresan, Yinghong Yu, Walter Hoersch, Thierry Pochet, and Angelo Auricchio," Optimizing Cardiac Resynchronization Therapy in Heart Failure Patients by Measuring Transient Changes in Sinus Rate During Pacing" Proceedings of the 23rd Annual EMBS International Conference, October 25-28, Istanbul, Turkey 2001.

[9] Virginie Le Rolle*, David Ojeda, and Alfredo I. Hernandez," Embedding a Cardiac Pulsatile Model into an Integrated Model of the Cardiovascular Regulation for Heart Failure Follow up", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 58, NO. 10, OCTOBER.

[10] Kyle Weigand1, Russell Witte1, Talal Moukabary4, Ike Chinyere, Jordan Lancaster1, Mary Kaye Pierce, Steven Goldman, Elizabeth Juneman," In-vivo Electrophysiological Study of Induced Ventricular Tachycardia in Intact Rat Model of Chronic Ischemic Heart Failure" IEEE Transactions on Biomedical Engineering.

[11] Natalia M. Arzeno, Mark T. Kearney, Dwain L. Eckberg, James Nolan, and Chi-Sang Poon," Heart Rate Chaos as a Mortality Predictor in Mild to Moderate

Heart Failure" Proceedings of the 29th Annual International Conference of the IEEE EMBS Cite International, Lyon, France August 23-26, 2007.

[12] Gaetano M. De Ferrari, Antonio Sanzo, Peter J. Schwartz," Chronic Vagal Stimulation in Patients with Congestive Heart Failure" 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.

[13] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M," Heart Disease Diagnosis Using Data Mining Technique" International Conference on Electronics, Communication, and Aerospace Technology ICECA 2017

[14] 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

