Model to predict heart failures through Deep Belief Network

Sanket Mahajan
dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

Pranav Madihalli
dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

Saurabh Mirajkar
dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

Aishwarya Jadhav
dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

Rubeena A Khan
dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

Mahesh Shinde
Dept. of Computer Engineering
Modern Education Society's College of Engineering
Pune - 01, India

***Abstract*—The decrease in the mortality rate across the planet has led an increase in the geriatric individuals. These individuals are really old and are more prone to certain ailments than the younger counter parts. This has been one of the biggest medical problems that have been straining the health sector considerably in the past few years. This is a problematic occurrence that has led to increased fatalities due to heart diseases. These cardiac ailments are extremely difficult to diagnose due to the inherent nature of these diseases which requires extensive tests and diagnosis. The delay in getting the results leads to the delay in providing prompt treatment to the patient which is highly necessary for heart diseases, failing which can lead to casualty. Therefore, an automatic approach for the diagnosis is needed to achieve prompt heart disease detection through the use of Deep learning methodologies. This research article describes a precise heart disease detection mechanism that utilizes K-means and Pearson Correlation along with Deep Belief Network and Decision Tree. The experimental results indicate an effective performance for the detection that is highly satisfactory.**

Keywords—Heart Disease Detection, K-Means Clustering, Deep Belief Network, Decision Tree.

I. INTRODUCTION

In recent years, heart disease has been one of the significant causes of death. This has had a lot of attention all over the world. The rise in the number of elderly people can be due to the substantial increase in the cases of heart disease in the population. The chances of heart attack and heart failure rise significantly as people age. The elderly are typically frail and have a weakened blood circulatory system, making them more vulnerable to heart disease and heart failure. As a result, the majority of heart disease patients are older people who need special attention and security.

Large advancements in medical treatments and services have been due to substantial studies in the medical model over the years. This ongoing study has been going on for decades, with researchers building on previous scientists' work. The incremental advancements led to the treatment of a variety of illnesses, both common and serious. Scientific medicine has resulted in a dramatic rise in human life expectancy as well as a substantial reduction in child mortality. This resulted in a vast number of older people and a slew of additional aging-related problems. Thanks to the lower life expectancy that prevailed at the time, these problems went relatively unnoticed.

As a result, much more research on this model is needed to allow a better approach to the prevention and treatment of various ailments. Hospitals have made substantial strides in the care of older patients, and better nursing has resulted in higher patient survival rates. Inadvertently, this increases the risk of diabetes, obesity, and cardiovascular disease by a factor of ten. The majority of treatments for heart disease have been focused on treatment; however, there is a need to incorporate a preventive strategy that can be used widely to reduce the risks associated with heart failure.

The heart disease appraisal remains one of the most antiquated methods used to analyze the likelihood that a certain symptom will result in a heart failure situation for the patient. To ensure that the diagnosis of heart failure disorder is correct, it must be confirmed. The heart disease theory is the most complicated methods because it is regulated by a broad range of human risk factors as well as a variation of the prevalence of various diseases such as vascular disease, Congestive Heart Disease, hypertension, and high blood pressure, both of which contribute significantly to heart failure. As a result, heart disease symptoms are very complicated and can occur unexpectedly at any moment. Therefore, a heart failure condition prediction for a patient is needed to better guarantee that the patient requires timely treatment. This is because heart failure is a very time-sensitive illness, and every second count when faced with such a situation. As a result, it is important that a predictive mechanism will forecast the real heart attack situation and it can better brace for the scenario, saving precious time and potentially saving a patient's life.

The prediction of heart attack scenarios will also assist doctors in determining whether or not a patient can be discharged. Because most elderly patients live alone, detecting any imbalances in their bodies when at home is daunting, and the time it takes to get the patient to a care provider for care following an episode could be particularly fatal for the patient, and the time could be the distinction between life and death. In this case, if a circumstance like this is expected, the doctors will not discharge the patient so that he or she stays under hospital supervision until the incident occurs to get prompt and accurate medication.

Machine learning is one of the most widely used methods for predicting heart failure. Machine Learning is a method that uses a set of historical data to forecast future events. The predictions are based on repeated trends in other data that enable the algorithms to recognize and expand on them to make accurate predictions. The algorithms are created specifically to detect patterns in data, which are then used to provide accurate predictions. Past data is required for the algorithms to function properly and infer potential events based on the insight provided.

Predicting cardiac failure necessitates a huge amount of data involving a variety of criteria as well as information about the patient's vital stats and other heart-related data. The data contains a vast number of attributes, and as the number of attributes grows, so does the prediction's accuracy. The Machine Learning model is a great alternative for this because it can make extremely precise predictions. This paradigm evaluates the risks and other metrics that can reliably measure the patient's situation and produce acceptable outcomes.

The heart attack failure prediction saves physicians and professionals a lot of time that would otherwise be spent reviewing multiple studies and the patients' past histories, which takes a long time. Furthermore, the characteristics and diseases that cause heart failure are many, and different attributes may have a significant impact on the outcomes. A vast array of characteristics adds to the prediction's difficulty, putting even more strain on the care practitioner in charge of the patient. The introduction of automated prediction by the use of Machine Learning is extremely feasible and a realistic solution to the platform due to the rise in difficulty along with the increasing workload on medical professionals.

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for future enhancement.

II. LITERATURE REVIEW

A. Naniwal et al., states that heart failure occurs due to a variety of factors that all contribute to the heart's deterioration. The majority of heart diseases affect elderly people, and there is a need to improve the diagnosis process to save lives. An ECG scan taken from the patient is used to make the primary diagnosis of Congestive Heart Failure [1]. The authors of this paper devise a method for detecting morphological changes in an ECG, which can be used to determine if the morphological changes are an appropriate representation of Congestive Heart Disease. The only disadvantage of this approach is that the authors only considered a limited range of features that result in sub-optimal results.

G. Valenza et al., introduces the Heart Rate Variability (HRV) spectral analysis as one of the most commonly applied methods for classification of the heart. However, there are several risks to this approach that must be addressed to make a correct diagnosis of Congestive Heart Disease. As a result, the authors have proposed a methodology for identifying the dynamics of a stable heart using the PAI or Parasympathetic Activity Index and the SAI or Sympathetic Activity Index [2]. The findings show that the proposed approach is very dependable. The only disadvantage of this approach is that the authors' analysis does not provide any evaluation of other applicable settings, such as physical exercise and other criteria.

S. Potturu et al. expands on the idea of using the nervous system in conjunction with the respiratory system to diagnose Congestive Heart Failure. The authors created a system of two states that employs a non-linear space model, with the variables being partial carbon dioxide and oxygen pressures in the alveoli. In the input control loop, the device identifies the delay, which is then used to identify based on the duration of the delay [3]. If the delay is important, it is an indication of Congestive Heart Failure. The experimental findings show that the suggested procedure provided a precise analysis. The key drawback of the suggested technique is that the authors did not use non-negativity theories to analyze this model.

A. Khayyat et al. address the medical disorder known as Congestive Heart Failure, which is characterized by slow fluid flow in the body, resulting in multiple swellings in the internal organs. This is a fatal illness that is commonly found in older patients and may have serious consequences. As a result, the authors propose a method for diagnosing this disease using a decision support method that can minimize or predict the chances of the patient being readmitted for the same condition (CHF) [4]. The efficacy was shown by the researchers by extensive experimentation, which obtained adequate results. The only disadvantage of this methodology is that the authors did not modify the decision support method to improve its precision with a More rigorous approach.

W. Pan et al. presents the paradigm of using a Heart Rate Variability index to diagnose Congestive Heart Failure disorder. According to the authors, the Autonomic Nervous System is an accurate indicator of a person suffering from Congestive Heart Failure. As a result, the authors devise a technique for identifying CHF using HRV and multi-frequency components Entropy using the HHT paradigm [5]. The experimental findings show that the new methods outperform the old ones in terms of precision. The main disadvantage of this approach is that the developers failed to account for the algorithm's shortcomings while developing the method, resulting in even lower accuracy.

E. Mbazumutima et al. sates there is a chance of predicting the myriad factors in an exercise procedure intended to determine the seriousness of the patient's Congestive Heart Failure condition. The researchers suggested an approach that employs Linear Regression (LR) to accurately estimate the patient's peak heart rate and oxygen levels. The findings of the experiments show that the suggested technique is highly reliable and can be used for estimation [6]. The methodology shows a reliable peak heart rate, but the estimates of peak oxygen intake are extremely unreliable, which is the paper's biggest drawback.

H. Wendt et al. introduces a variety of indices that have been developed to systematically classify a vast range of Heart Rate Variability variables and their temporal dynamics. The authors developed a method that uses non-Gaussian multi-scale representations on wavelet p leaders and the Apriori system to distinguish between a stable heart and a heart that is suffering from Congestive Heart Failure [7]. For its success measurement, the system was evaluated and tested, and the results showed that it performed well. The writers have not used the various variations of higher-order cumulants in the system, which will improve the output metrics, which is the key disadvantage of this approach.

B. Hu et al. expands on the subject of identifying Congestive Heart Failure in patients, claiming that diagnosis with relative precision necessitates the use of ECG or Electrocardiogram-based Heart Rate Variability tests. The authors propose a method for incorporating several time scales into patient interpretation to improve prediction accuracy. The suggested methodology includes time-based features for study,

The proposed system for attaining heart disease detection through the use of Deep belief Networks has been described in the system overview diagram above. The procedural step utilized to achieve the system has been elaborated in the section below. and the experimental findings show that it achieves a high degree of precision as compared to other techniques [8]. The biggest disadvantage of this paper is that it has a greater time complexity than most conventional approaches.

A. Windmon et al. proposes the idea of a smartphone application that can be used to detect a rise in cough in patients with congestive heart failure since cough is one of the first signs of chronic obstructive pulmonary disease, which is commonly exacerbated by CHF. TussisWatch is a tool developed by the writers that are used to distinguish cough episodes and correlate them with other outcomes for the diagnosis of CHF [9]. In contrast to popular methods, the experimental findings show a significant improvement in CHF detection. The biggest disadvantage of this method is that the classification accuracy has not changed, resulting in irregular performance.

D. Destiani et al. presents that Congestive Heart Failure is a heart defect that is caused by defects in the heart muscle and other core muscle disorders. An ECG scan of the heart and its beats can be used to detect congestive heart failure. Using the Polak-Ribiere conjugate and Artificial Neural Networks on the Discrete Wavelet Decomposition, the authors formulate a strategy for predicting Congestive Heart Failure [10]. The suggested method has been thoroughly tested and achieves excellent results. The only disadvantage of this approach is the system's increased computational complexity.

M. Alex et al. address the diagnostic and prognostic paradigms for different heart-related disorders. In recent years, the number of deaths caused by heart disease has increased. This is because diagnosing the disease takes time because by the time it is diagnosed, the harm has already been done. As a result, the authors formulate a strategy for predicting multiple cardiac diseases using Artificial Neural Networks and K means for precise heart attack forecasts at an early stage, which may help save many lives on time [11]. The writers used a Data Mining technique with a high space complexity, which is the only disadvantage.

A. Massaro et al. states that a vast number of smart health networks have risen in popularity in recent years. These systems are capable of making a vast number of observations based on data gathered from a variety of sources. The authors developed a method for using a Multilayer Perceptron in conjunction with Artificial Neural Networks to accurately model Congestive Heart Failure conditions in patients before their onset [12]. This makes for a quick diagnosis that will help relieve the patient's discomfort, as well as giving physicians more time to get accurate care at an early stage.

 ISSN: 2582-3930

III. PROPOSED METHODOLOGY

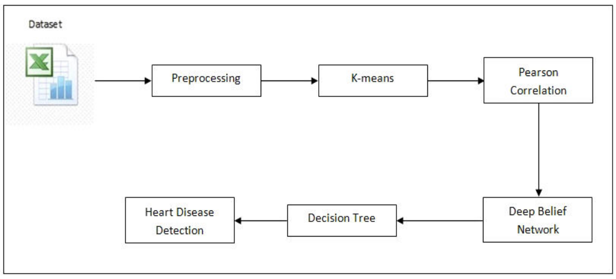

Figure 1: Proposed model System Overview

Step1: Dataset collection, preprocessing and Labeling – The proposed system performs the heart disease detection based on the input provided in the form of a dataset. This dataset is downloaded in the form of a workbook format from the URL - https://figshare.com/s/3ea30ba1ac39af387358 and provided to the system as an input.

The downloaded dataset from the above URL consists of a variety of heart related data for a number of patients. The dataset consists of parameters that can be highly useful in the realization of the heart disease detection system designed in this research article. This dataset consists of attributes such as, patient ID, RR segment intervals in ms, QT segment interval in ms, QTc, PR segment interval in ms, QRS segment interval in ms, P wave axis, QRS axis, T wave axis, Adjusted Charlson Comorbidity Index (ACCI), ecgdept which consists of elements such as – E = Emergency, H = Health examination, O = Outpatient, I = Inpatient, and Ecgsource.

These attributes provide extensive information about the Electrocardiogram of the patients which is one of the most effective methods to determine the presence of heart disease. Out of all the values and the attributes provided in the dataset, only the relevant attributes that are required by our approach have been selected for the evaluation. These values are the RR intervals, PR intervals, QRS interval and the QT intervals. These attributes are extracted into a double dimension list and preprocessed for the utilization as an input to the next module of the system for further processing.

Step 2: K-means Clustering – The preprocessed and labeled list obtained in the previous step is utilized as an input in this step of the procedure. This step of the heart disease procedure performs the K Means clustering for the segregation of the data into semantic groups to perform accurate prediction of the heart disease in an individual. This procedure is performed through the steps given below.

Distance Evaluation – The evaluation of the distance in the attributes is achieved by the use of the Euclidean Distance of the selected attributes that are stored in the form of rows in the list. These attributes are the QT, QRS, PR and RR interval values in ms. The Euclidean distance is measured for each of the rows and the average of the achieved distance is added to the end of the respective row. This distance is the Average row distance. The equation used for the calculation is given in the equation 1 given below.

$$ED = \sqrt{\Sigma (A_{Ti} - A_{Tj})^2}$$

Where,

ED=Euclidian Distance

A_{Ti}=Attribute at index i

A_{Tj}= Attribute at index j __________ (1)

Centroid Estimation – The Row distances achieved and added to the end of the respective rows is utilized in this step for the purpose of sorting the entire list into an ascending order. Once the sorted list is achieved, a number of Data points are aggregated at randomly. These are K number of indices that are randomly fetched as the data points. These data points in turn are used to grab the row distance of the particular data point which is now referred to as the centroid. This turns into the boundary for the fetched row distance.

The grabbed average distance and the row distance are considered as the boundaries of the cluster that will be achieved by the addition and subtraction of these values from one another. These achieved boundaries will be useful in the next step for the assimilation of the clusters.

Cluster Formation – The row distances in the list are now subjected to the K boundaries attained in the previous step for the purpose of achieving the clusters. These clusters are based upon the input values of the PR, RR, QRS and QT intervals which will be provided for the examination of the correlation in the next step of the system.

Step 3: Pearson Correlation – This step of the procedure utilizes the cluster obtained in the preceding step for the evaluation of the correlation between the user input and the clusters generated by the system on the input dataset. The Pearson correlation approach is used to achieve the correlation which utilizes the user input for the attributes as the value of x list and the cluster attributes as the y list values. These values are subjected to the equation 2 given below for the calculation of the correlation value.

This process of Pearson Correlation can be depicted using the below mentioned algorithm 1.

ALGORITHM 1: Pearson Correlation Estimation for each of the Clusters

---

//Input : K-NN Cluster List KNC_L

// Input : User attribute list X[]

```
//Output:Correlation List PCR_L
pearsonCorrelationEstimation(KNC_L)
Start
CR_L =Ø
  for i=0 to Size of KNC_L
    S_GCL= KNC_L [i] [S_GCL = Single Cluster]
  for j=0 to Size of S_GCL
        ROW = S_GCL [j]
        Y[]→ROW
        PR_C=pearsonCorrelation(X,Y)
        ROW= ROW+PR_C
         S_GCL = S_GCL +ROW
  end for
     PCR_L = PCR_L + S_GCL
  end for
   return PCR_L
Stop
```

The clusters achieved in the previous step have been utilized as an input in this step of the approach. This is combined with the user input provided as the value of x which is then used to correlate with the cluster values. The Pearson correlation utilizes the clusters and extracts the individual row values for the calculation of the correlation with each of the rows with respect to the user input. This is performed through the use of the equations illustrated below.

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(x^2 - \frac{\sum x^2}{n}\right)}\sqrt{\left(y^2 - \frac{\sum y^2}{n}\right)}} \quad \_\_\_\_\_\_\_\_ \ (4)$$

Where,
x is the user input attributes
y is the attributes extracted from the dataset
n is the total number of entries
r= correlation value in between -1 to +1.

The correlation is achieved for each of the rows present in the cluster. This is repeated for all the clusters and their respective rows for the user input values. The correlation values of all of the rows of the cluster are averaged to get a Pearson correlation value of the entire cluster. The achieved values of correlation are stored in the form of a list. These correlation values are in the range of 1 to -1. The cluster list is then sorted in the descending order of the correlation values. The top 3 clusters from the list are then selected and provided to the Deep Belief Network in the next step to achieve the heart disease detection.

Step 4: Deep Belief Network – The top 3 clusters with the maximum amount of correlation with the user input are provided to this Deep belief network Module as an input. The Deep Belief Network is tasked with determination of the hidden layer values through the Restricted Boltzmann machines to achieve the output layer values. 21 random weights are assigned to the input layers with a bias set at 1. The sigmoid activation function is used along with the contrast divergence and linear summation of the input. This can be mathematically depicted in the equations 3 and 4 given below.

$$T = \left(\sum_{K=0}^{n} A_T * W\right) + B \quad \_\_\_\_ \ (3)$$

$$H_{LV} = 2\left(\left(1/\left(1+\exp^{\wedge}(-T)\right)\right)*2T\right) - 1 \quad \_\_\_\_\_\_ (4)$$

Where,
n- Number of attributes
A_T- Attribute Values
W- Random Weight
B- Bias Weight
H_{LV} – Hidden Layer Value

The 3 Hidden layers then utilize the ReLU activation function to determine the DBN probability. These probability values are then stored in the form of a list. This list is then sorted in a descending order to achieve the highest probability values on top of the list. The higher probability indicates a more precise prediction with respect to the user input. This list needs to effectively classify in the next step for achieving the results of the heart disease detection.

Step 5: Decision Tree – The attained probability scores from the Deep belief network execution in the previous step are being utilized as an input for classification purpose. The Decision Tree approach provides precise segregation of the relevant probability scores for an accurate output. The If-then rules provide effective realization of the classification of the probability scores. The implementation of the Decision tree also serves as an improvement in the Heart Disease Detection Methodology by eliminating any false positives that inadvertently creep into the output. The detection of classified heart disease is then displayed to the user.

IV. RESULT AND DISCUSSIONS

The proposed system for the detection of the Heart Diseases has been developed using Python programming language. The Spyder IDE has been used to attain the development of the technique on a laptop. This laptop was powered by an Intel i5 CPU with a combination of 4GB RAM and 500GB Hard drive storage.

The performance of the prescribed detection mechanism for the Heart Diseases needs to be evaluated to determine efficiency of the approach. The evaluation has been performed for the error achieved through the execution of the proposed methodology. The error is determined by the implementation of the RMSE or Root Mean Square Error.

 ISSN: 2582-3930

The error of the detection is useful in understanding the overall accuracy of the prescribed system.

The RMSE approach utilizes a pair of continuous and correlated variables for the assessment of the error achieved between these entities. The variables selected for the assessment of our methodology are the expected heart disease detections and the obtained heart disease detections. This has been illustrated mathematically in the equation 5 given below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^2}{n}} \quad \text{\_\_\_\_\_\_ (5)}$$

Where,

- Summation $(x_1 - x_2)^2$ - Differences Squared for the summation in between the expected heart disease detections and the obtained heart disease detections

n - Number of samples or Trails

Extensive evaluation has been performed and the outcomes have been tabulated in the table given below.

Experiment No	No of Expected Heart Disease Detections	No of Obtained Heart Disease Detections	MSE
1	12	10	4
2	10	8	4
3	13	9	16
4	8	7	1
5	7	5	4

Table 1: Mean Square Error measurement

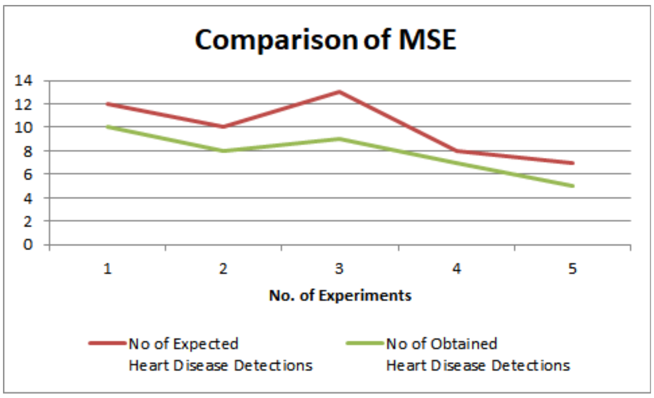

Figure 2: Comparison of MSE in between No of Expected heart disease detections V/s No of obtained heart disease detections

The values recorded in the table have been used to achieve the line graph plotted in the figure 2 above. These values have been achieved for 5 trials of experimentation conducted for the heart disease detection with varying user inputs. The error attained between the expected heart disease detection and the achieved heart disease detection has been calculated for each of the trials. The evaluation outcomes have attained a value of MSE and RMSE of 5.8 and 2.4 respectively. This indicates a respectable performance of the detection approach which is well within the limits. The experimentation has been considerable in the extraction of the performance metrics which have achieved satisfactory outcomes.

V. CONCLUSION AND FUTURE SCOPE

The presented system for the heart disease detection has been illustrated in this research article. There have been a considerable increase in fatalities have been caused due to heart diseases. The cardiac ailments have been extremely problematic to deal with due to the large amount of time and resources required to accurately diagnose the disease. These diseases have been effectively useful in determination of the heart diseases. This research article achieves the determination of the heart disease through the use of a dataset containing parameters of Electrocardiogram. This dataset is preprocessed, labeled and provided to the K-Means for the purpose of achieving the clusters. The K-Means approach utilizes distance evaluation and centroid formation to achieve the clusters. These clusters are provided to the Pearson Correlation for the purpose of determining the correlation between the user input and the clusters. The obtained clusters have been provided to the Deep Belief Networks for the evaluation of the hidden layer and the output layer. The resultant values are then provided to the Decision Tree approach to achieve the classification of the heart disease detection. The experimental evaluation has resulted a highly satisfactory performance.

The future research approach can be made in the direction of improving the heart disease detection system into an API for easier integration into existing systems and also this can be deployed on huge dataset in terms of GB to learn more efficiently by using cloud infrastructure.

REFERENCES

[1] A. Naniwal, Y. Kumar and B. Jha, “Morphological Changes in Congestive Heart Failure ECG”, 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA), 2016.

[2] G. Valenza et al, “ECG-Derived Sympathetic and Parasympathetic Nervous System Dynamics: A Congestive Heart Failure Study”, IEEE Computing in Cardiology, 2018.

[3] S. Potturu, T. Pal and S. Maka, “Modelling and Analysis of Respiratory System with Congestive Heart Failure”, Fifth Indian Control Conference (ICC), 2019.

[4] A. Khayat et al, “Decision Support Tool to Estimate and Reduce the Probability of Readmission for Congestive Heart Failure Patients”, IEEE Access, 2019.

[5] W. Pan et al, "Multi-Frequency Components Entropy as Novel Heart Rate Variability Indices in Congestive Heart Failure Assessment", IEEE Access, 2019.

[6] E. Mbazumutima et al, "A Study of the Possibility of Predicting Main Parameters of a Cardiopulmonary Exercise Testing in Patients with Congestive Heart Failure", IEEE Access, 2019.

[7] H. Wendt, P. Arby et al, "Wavelet p-Leader Non-Gaussian Multiscale Expansions for Heart Rate Variability Analysis in Congestive Heart Failure Patients", IEEE Transactions on Biomedical Engineering, 2018.

[8] B. Hu et al, "Multiple Time Scales Analysis for Identifying Congestive Heart Failure Based on Heart Rate Variability", IEEE Access, 2019.

[9] A. Windmon et al, "TussisWatch: A Smart-phone System to Identify Cough Episodes as Early Symptoms of Chronic Obstructive Pulmonary Disease and Congestive Heart Failure", IEEE Journal of Biomedical and Health Informatics, 2018.

[10] D. Destiani, Adiwijaya and D. Utama, "Study of Wavelet and Line Search Techniques on Modified Backpropagation Polak-Ribiere Algorithm for Heart Failure Detection", 6th International Conference on Information and Communication Technology (ICoICT), 2018.

[11] M. Alex et al, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication and Signal Processing, April 4-6, 2019.

[12] A. Massaro et al, "Neural Networks for Automated Smart Health Platforms oriented on Heart Predictive Diagnostic Big Data Systems", AEIT International Annual Conference, 2018.