Monitoring Health Records For Chronic Disease Prediction And Risk Stratification

Nehal N Ghosalkar
Department Of Computer Engineering
Maharashtra, India
Email: nehal.ghosalkar@spit.ac.in

Kailas Devadkar
Department Of Information Technology
Maharashtra, India
Email: kailas_devadkar@spit.ac.in

***Abstract*—The chronic diseases related to heart are the primary purpose behind countless deaths over the last couple of decades and has developed as the most dangerous illness, in India as well as in the entire world. In this way, there is a need of dependable, precise and reliable framework to analyze such ailments in time for appropriate treatment.The ongoing advances in innovation have encouraged the standard gathering and storage of medical information that can be utilized to help medicinal choices. Be that as it may, in many nations, there is a requirement to gather patient's information in digitized structure. At that point, the gathered information are to be examined all together for a therapeutic choice to be made, regardless of whether it includes prediction of disease, its diagnosis or course of treatment. In this paper,dataset from UCI archive is utilized for Heart ailment analysis. The right finding execution of the programmed conclusion framework is assessed by utilizing characterization exactness, affectability and particularity examination. The investigation demonstrates that, the SVM with CNN-MDRP calculation have better decision for therapeutic infection analysis application.Early recognition of cardiovascular ailments and constant supervision of clinicians can lessen the death rate. A precision dimension of 96.77 percent exactness was found from the proposed framework.**

***Index Terms*—Data Mining, Machine Learning, Healthcare, CNN-MDRP, Chronic Heart Disease (CHD), Linear Support Vector Machine (LSVM)**

I. Introduction

The use of machine learning techniques in medical field is a subject of good investigation, that principally focuses on displaying some of the human activities or thinking forms and perceiving certain diseases from a scope of input sources. Diverse application zones are information revelation [10] and medicinal strength frameworks, that grasp genetic science and DNA analysis. The Data Mining strategies can be used to bring down the mortality rate, to improve the accuracy in disease prediction and also primarily reducing its diagnosis time.[3]
Human faces many issues associated with the chronic diseases the common reason behind its increase are improper living habits, deficient exercise, unhealthy diet, and irregular sleeping [3]. Eightieth of individuals around the world, spend additional quantity on the diagnosis of chronic sickness [1]. Individuals offers additional aid for correct prediction of sickness [1]. In several regions, different diseases are caused due to the environmental factors and every individuals lifestyle [1]. Sometimes it may result in the incorrect decision concerning the disease prediction. Because of preliminary disease prediction, it will cut back the danger of sickness and patient gets diagnosed as early as possible.
In a diagnosis drawback, what's required might be a lot of examples or characteristics that are illustrative of the considerable number of varieties of the diseases. The models must be picked precisely if the framework is to perform loyally and quickly. The very reality that there's no compelling reason to give a selected algorithm on an approach to recognize the ailment, shows a genuine preferred standpoint over the applying of Machine Learning methodologies to the present kind of issues.
For instance, amid a chronic disease detection task, it's vital for the prediction on healthy patients to be precised as high as possible, as a misclassification amid this class may prompt a healthy patient undergoing treatment for certain illness for reasons unknown. CHD have risen on the grounds that the perfect executioner in each urban and country territories in the vast majority of the nations. It's determined that, in a few cases because of wrong prediction it has resulted in patient's health compromise. In the vast majority of the developing nations specialists aren't wide reachable for the correct diagnosis. Thus, such machine-driven framework will encourage to medicinal calling to help specialist for the right analysis well before. Different researchs about have been led to improve the accuracy of disease classifciation from a vast information.
The existing work has only considered about just structured information. For unstructured information, convolutional neural system (CNN) is utilized to remove content qualities naturally. Structured information is widely used for the Chronic disease prediction apart from unstructured data. However by the employment of a convolutional neural network, it becomes straightforward to accommodate unstructured knowledge additionally [1]. The convolutional neural network is deep learning algorithmic rule that extracts the options mechanically from the big dataset and gets the correct result [1].
To reduce the incorrect prediction of disease we'll get to access the organized and unorganized data in healthcare eld to survey the existence of the disease. At last, we use Convolutinal neural network -based multimodal disease risk prediciton (CNN-MDRP) algorithmic program for organized and text data. The sickness hazard display is gotten by the blend of

organized and unstructured text choices. Through the trial, we tend to make a determination that the execution of CNN-MDRP is best than various existing ways.

The rest flow of the paper is sorted out as follows: Section II quickly audits some previously proposed strategies in Heart disease diagnosis. Area III depicts the process for coronary illness finding; Section IV highlights the methodology utilized for CHD analysis. Also, the proposed CNN-MDRP calculation and the systems utilized are talked about nitty gritty in Section IV. The experimental results are given in Section V. Segment VI finally concludes the paper.

II. Background

Theodora Brisimi, et.al (2018) presented in this paper [15], two new strategies: K - LRT, which is a probability based proportion technique , and a joint clustering and classication (JCC) strategy it recognizes shrouded quiet groups and adjusts classiers to each cluster. The prediction problem is formulated as binary classication problem and think about an assortment of ML techniques, including support vector machines (SVMs), logistic regression, and decision tree. To balance a harmony among exactness and interpretability of the forecast, or, in other words a restorative setting, they did this predictions based on the patients medical history.

Akhilesh Kumar Yadav, et.al (2013) presented in this paper [8], that dierent analytic tool has been used to extract information from huge datasets such as in medical eld where a huge data is available. The classication becomes inecient due to noise, high dimensional and missing values. Due to the dierent challenges have to face while performing data analytics clustering is used in replace of it. The foggy k-mean clustering based novel technique need to be developed is the main focus of authors. The proposed algorithm has been tested by performing dierent experiments on it that gives excellent result on real data sets. In real world problem enhanced results are achieved using proposed algorithm as compared to existing simple k-means clustering algorithm.

Min Chen, et.al (2017) proposed in this paper [7], a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. In order to make predictions related to the chronic disease that had been spread within several regions, various machine learning algorithms were streamlined here. A latent factor model was utilized to reconstruct the incomplete type of data present within the gathered data. A chronic disease of cerebral infarction was utilized in order to perform various experiments to predict the accuracy of proposed method. 94.8 percent of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

Sanjay Chakrabotry et.al, (2014) stated in this paper [9], that powerful tool clustering is used as dierent forecasting tools. The weather forecasting has been performed using proposed incremental K-mean clustering generic methodology. The weather category has been dened in dierent clusters and a new data is checked by incremental K means to group it into existing clusters. The used data set contains the weather forecasting information of west Bengal that is able to reduce the air pollutions consequences. The weather events forecasting and prediction becomes easy using modeled computations. In the last the authors have performed dierent experiments to check the proposed approach correctness.

Chew Li S. et.al, (2013) presented in this paper [10] particular university student results has been recorded to keep a track using Student Performance Analysis System (SPAS). The design and analysis has been performed to predict students performance using proposed project on their results data. The data mining technique generated rules that are used by proposed system to gives enhanced results in predicting student performance. The students grades are used to classy existing student using classication by data mining technique.

Qasem A. et.al, (2013) presented in this paper [11] that data analysis prediction is considered as import subject for forecasting stock return. The data analysis future can be predicted through past investigation. The past historical knowledge of experiments has been used by stock market investors to predict better timing to buy or sell stocks. There are dierent available data mining techniques out of all a decision tree classier has been used by authors in this work.

K.Rajalakshmi et.al, (2015) presented in this paper [12] a study related to medical fast growing eld authors. In this eld every single day a large amount of data has been generated and to handle this much of large amount of data is not an easy task. The medical line prediction based systems optimum results are produced by medical data mining. The K-means algorithm has been used to analyze dierent existing diseases. The cost eectiveness and human eects has been reduced using proposed prediction system based data mining.

Bala Sundar V et.al, (2012) examined in this paper [13] real and articial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique results to check its accuracy. The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis and each cluster has its outputs with nearest mean. The cluster centroid and Euclidean distance formula has been used between data to perform the above mentioned task. The proposed scheme of integration of clustering has been tested and its results show that highest robustness, accuracy rate can be achieved using it.

Daljit Kaur et.al (2013) explained in this paper explained [14] that data contained similar objects has been divided using clustering. The motive of authors is to reduce its drawback to make it more eective and ecient. The proposed algorithm has been tested and results shows that it is able to reduce eorts of numerical calculation, complexity along with maintaining an easiness of its implementation. The proposed algorithm is also able to solve dead unit problem.

As different research are done by authors using various Machine Learning Algorithms, it is seen than predicting chronic diseases is a complex study.

III. Methodology and Data Analysis

In this study, an effective machine learning algorithm was looked over some accessible algorithms in order to recognize the probability of having heart disease from a huge dataset. The well ordered structure methodologies of the proposed framework and the work process of the total framework have been referenced below The Design is isolated into three fundamental stages: Initial, Middle and Last stage. The Initial stage is identied with Data aggregation and Analysis. The next stage includes distinctive substages like Feature Selection, Training the SVM Model and Measure the slipup estimations. Last stage incorporates the Visualization of the data. [3]

Data Imputation:

For patients examination information, there'll be an oversized range of missing information because of human error. Along these lines, we will fill the organized data. Prior to data ascription, we tend to at first decide inadequate medical data so as to alter or erase them to upgrade the data quality. At that point, we tend to utilize data incorporation for information pre-preparing. we will incorporate the healthcare data to guarantee data atomicity. For data attribution, we tend to utilize the latent factor model that is given to illustrate the recognizable factors as far as the inert factors.

Splitting the data into Training and Testing Set

The inmate dataset information contains structured and unorganized data. The structured information includes the data from the laboratory and therefore the patients basic data like the patients age, their gender, previous medications and lifestyle etc. whereas the unorganized text information consists of the patients narration of his/her health problem, the doctors interrogation records and diagnosing etc. The main aim of this experiment is to predict whether or not a patient has the possibilities of getting any sort of diseases from his medical record. During this we have a tendency to divide the information into coaching data and take a look at data. For S-data, we have a tendency to use standard machine learning algorithmic program, i.e., Linear SVM algorithmic program to predict the chance of sickness and For Unstructured information, we have a tendency to predict the chance of sickness by the utilization of CNN-MDRP.

Further the Model will be Classied as Structured and Unstructured Data:

The predictive analysing technique is the technology which can predict the future possibilities from the existing data. The data which is similar can be clustered in one cluster and another in the second cluster. The clustered data will be given as input for the classication in which SVM classier is used to classify data. The Convolutional Neural Network algorithm will take attribute number and instance value as input and give result in the form of relationship between the attributes. In the last step, the SVM classier will be applied which can classify data into two classes. The first class will be of the instances which have regional disease and second class is of an instance which does not have regional disease.

No.	Attribute	Description
1	Age	Age in years
2	Sex	Male = 1, female = 0
3	Cp	Chest pain type (typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic = 4)
4	Trestbps	Resting blood sugar (in mm Hg in case of admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl (true = 1, false = 0)
7	Restecg	Resting electrocardiographic results (normal = 0, having ST-T wave abnormality = 1, left ventricular hypertrophy = 2)
8	Thalach	Maximum heart rate
9	Exang	Exercise-induced angina
10	Old peak	ST depression induced by exercise comparative to rest
11	Slope	Slope of the peak exercise ST segment (upsloping = 1, flat = 2, down sloping = 3)
12	Ca	Number of major vessels which are colored by fluoroscopy
13	Thal	Normal = 0, fixed defect = 2, reversible defect = 3

Fig. 1. Selected Heart Disease Attributes

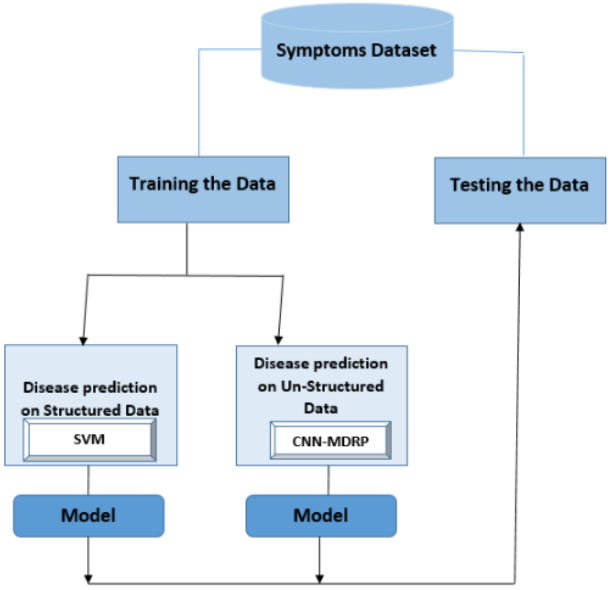

Fig. 2. Data Flow Model

IV. Proposed System Architecture

In the proposed framework we would first be able to get the extensive volume of a enormous information, at that point that information is considered as training information. SVM calculation is utilized for the classification of the information. At that point after the elucidation the emergency clinic information comparable sort of information can be put away. At that point CNN extricate the content attributes naturally. In that we utilize a CNN MDRP calculation that utilizes both

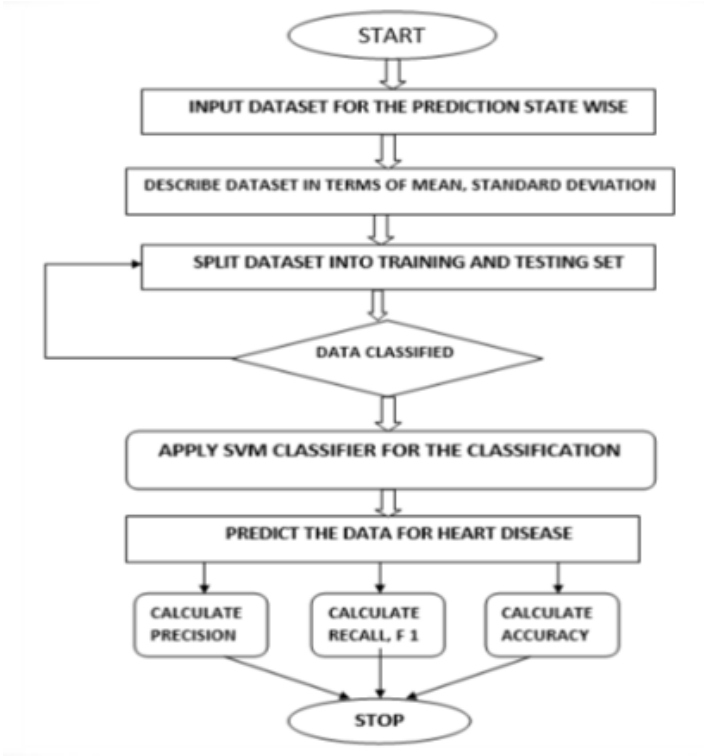

Fig. 3. System Architecture

structured and unstructured clinic information by choosing the characteristics automatically from a huge chunk of data. This improves the sickness expectation instead of recently chosen attributes. CNN-MDRP calculation improves the accuracy of the consequence of an ailment forecast over an expansive volume of information from medical clinics.

Algorithm used: CNN-MDRP

CNN-UDRP (Conventional neural network unimodal disease risk prediction) just uses the content information to predict whether the individual has illness or not. With respect to both organized and unstructured content information, we structure a CNN-MDRP (Conventional neural network unimodal disease risk prediction) calculation dependent on CNN-UDRP. The handling of content information is comparative with CNN-UDRP and calculation techniques are likewise comparable with CNN-UDRP calculation.

Stage 1: Select the particular preparing parameters.

Stage 2: We utilize stochastic gradient strategy to prepare parameters, lastly reach the risk stratification of whether the individual suffers from the disease or not.

Algorithm used: LinearSVM

An SVM is associate degree economical binary classifier The LinearSVM algorithmic program seeks a separating hyperplane within the feature area, so data focuses from the

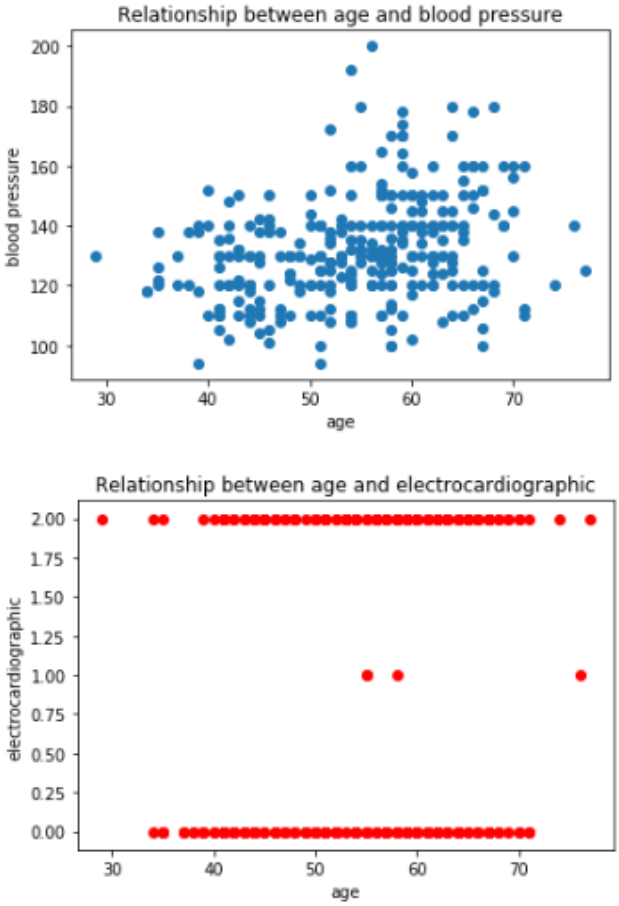

Fig. 4. Confusion Matrix

Fig. 5. Confusion Matrix

2 totally different classifications dwell on totally unique sides of that hyperplane. We figure the gap of every computer document from the hyperplane. The base over of these separations is named margin. The objective of SVMs is to search out the hyperplane that has the most margin.

Support Vector Machine (SVM) is a class of all inclusive feed forward system. SVM can be utilized for example characterization and nonlinear relapse. All the more absolutely, the SVM is an inexact usage of the technique for basic disease minimization. This guideline depends on the reality the error rate of a learning machine on test information is limited by the entirety of the preparation blunder rate. The SVM can give great speculation execution on example order problem.

Ideal Hyperplane for examples : Consider a preparation test where xi is called information design for the ith occurrence and yi is the relating target yield. With example spoken to by the subset yi= +1 and the example spoken to by the subset yi= - 1 are straightly divisible. The condition as a hyperplane that does the partition is:

$$W^{\mathrm{T}}x \ + b = 0(1) \tag{1}$$

where x = information vector and w = movable weight vector, and b = predisposition. Hence,

$$W^{\mathrm{T}} + X_{\mathrm{i}} + b \geq 0 for y_{\mathrm{i}} = +1 \tag{2}$$

$$W^{\mathrm{T}} + X_{\mathrm{i}} + b \geq 0 for y_{\mathrm{i}} = -1 \tag{3}$$

For a given weight vector w and a bias b, the separation between the hyperplane defined in above eq.1 and closest data point is known as the margin of separation.

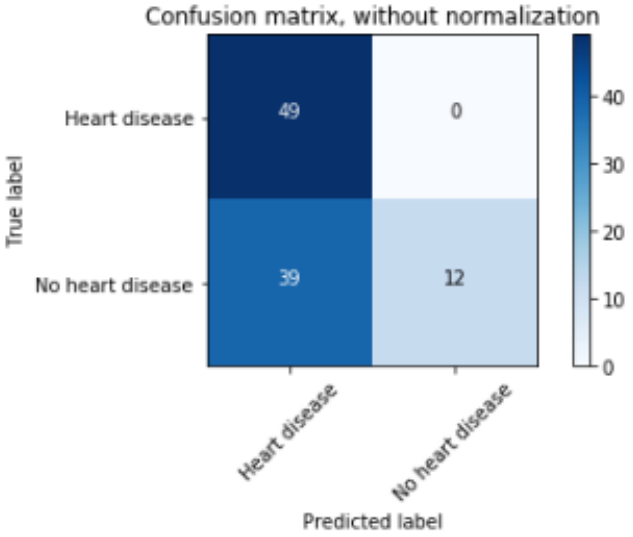

Fig. 6. Confusion Matrix

V. Results

At first, sample size, network size, test measure, selection of model and feature extraction and classification are considered to be the key parameters for the investigation of a reasonable system configuration issues identified with learning and speculation. While experimenting it is watched that, right and complete information gathering strategy is the correct course for the choice of best classifier. For assessing speculation execution as for precision, affectability, and explicitness dataset is parceled into number of subsets (i.e training set and testing set)

The experiment conducted on the training data consisting of medical symptoms of patients. The system uses a Linear SVM algorithm along with CNN-MDRP for prediction of disease based on their patient symptoms. We have achieved the performance and accuracy of Linear SVM algorithm as 96.77 percent which is accurate for prediction of symptoms from the diseases. The execution time is 0.128 seconds. As well as the performance of disease prediction with the help of unstructured data has increased compared from the previous work done.

Notwithstanding the previously mentioned assessment criteria, we use use receiver operating characteristic (ROC) curve and the area under curve (AUC) to assess the upsides and downsides of the classier. The ROC bend demonstrates the exchange off between the true positive rate (TPR) and the False positive rate (FPR), In the event that the ROC bend is nearer to the upper left corner of the chart, the model is better. The AUC is the region under the bend. At the point when the territory is more like 1, the model is better.[7]

In our proposed model we achieved the AUC as 0.8317 which suggests it a better model and the ROC curve is also upto the mark.

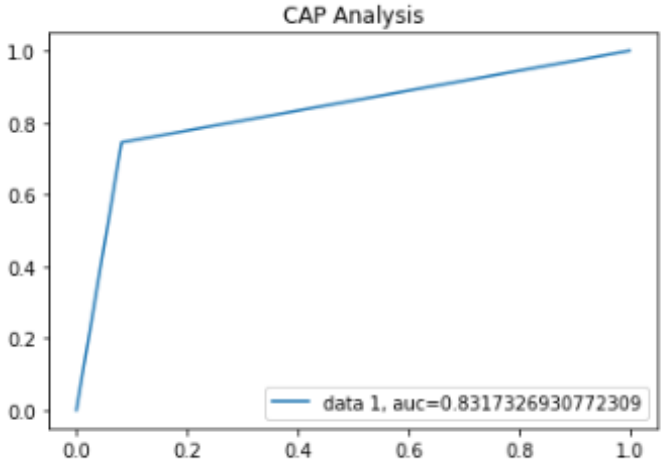

Fig. 7. ROC Curve

VI. Conclusion

Chronic Disease diagnosis has turned out to be exceptionally credited with the advancement of innovation recently. Moreover the Data mining and specialized apparatuses have improved the medicinal practice execution to a more noteworthy degree. Here we have proposed a Convolutional Neural Network System for the conclusion of CHD by method of Linear Support Vector Machine. Accordingly the conclusion of

Heart ailment is done using distinctive information tests from assorted patients and the outcomes have meant that SVM with Linear kernel is great in the analysis of Heart related illness. The characterization accuracy, affectability, and particularity of the LSVM have been observed to be high in this manner making it a decent alternative for such types of finding. Thus, in this paper, we leverage not only the structured data but also the unorganized text data of patients based on the proposed CNN-MDPR calculation. We find that by joining these two information, the exactness rate achieved is 96.77 percent, in order to better the prediction risk of CHD.

This framework leads in low time utilization and negligible cost workable for illness forecast. In future work, we may include more ailment into it so the general public gets more advantages about this framework.

REFERENCES

[1] Abdelghani Bellaachia and Erhan Guven *Predicting Breast Cancer Survivability Using Data Mining Techniques*.Washington DC 20052, vol. 6,pp. 234-239, 2010.

[2] Oyelade, O. J, Oladipupo, O. O and Obagbuwa *Application of k-Means Clustering algorithm for prediction of Students Academic Performance*. International Journal of Computer Science and Information Security, I.C, vol. 7, pp. 123-128,2010.

[3] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012) *Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity* . Middle-East Journal of Scientic Research,, vol. 12,pp. 959-963 2012.

[4] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S *Reducing the Time Requirement of K-Means Algorithm*. IEEE Transactions on Power Electronics, PLoS ONE, vol. 7, pp-56-62, 2012.

[5] Pranjul Yadav, Michael Steinbach, Vipin Kumar and Gyorgy Simon *Mining Electronic Health Records (EHRs): A Survey*. ACM Computer Survey, Article 85, 2018.

[6] Kajal C. Agrawal and Meghana Nagori *Clusters of Ayurvedic Medicines Using Improved K-means Algorithm*. International Conf. on Advances in Computer Science and Electronics Engineering, ,volume 23, 2013

[7] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017) *Disease Prediction by Machine Learning over Big Data from Healthcare Communities*. IEEE Transactions, vol. 15, pp- 215-227, 2017

[8] Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal, *Clustering of Lung Cancer Data Using Foggy K-Means*. International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, pp.121-126, 2013.

[9] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey, *Weather Forecasting using Incremental K-means Clustering*, Proceedings of the IEEE Region 10 Conference, vol. 8,2014.

[10] Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, *Student performance analysis system (SPAS)*. Information and Communication Technology for The Muslim World (ICT4M), 2014.

[11] Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi, *Predicting Stock Prices Using Data Mining Techniques*, The International Arab Conference on Information Technology ACIT2013, vol. 23, 2013.

[12] K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), *Comparative Analysis of K-Means Algorithm in Disease Prediction* International Journal of Science, Engineering and Technology Research (IJSETR) , Vol. 4, 2015.

[13] Bala Sundar V, T Devi and N Saravan, *Development of a Data Clustering Algorithm for Predicting Heart*, International Journal of Computer Applications, vol. 48, 2012.

[14] Daljit Kaur and Kiran Jyot, *Enhancement in the Performance of K-means Algorithm*, International Journal of Computer Science and Communication Engineering, vol. 2, 2013.

[15] Theodora S. Brisimi, TingTing Xu, Taiyao Wang, Wuyang dai, William g. adamS, IEEE transaction, 2018. *,Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classication Approach*, IEEE transaction, 2018.

[16] Po-Yen Wu, Chih-Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Homan, May D. Wang *Omic and Electronic Health Record Big Data Analytics for Precision Medicine.* IEEE Transactions on Biomedical Engineering, Volume: 64, 2017.

[17] Robert D. and Patricia E. Kern Stud, *Using EHRs and Machine Learning for Heart Failure Survival Analysis.* Health Technology Inform Manuscript, 2016.

[18] Singh Navdeep, Jindal Sonika, *Heart disease prediction using classication and feature selection technique* International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4, 2018.