International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 07 | July -2019 ISSN: 2590-1892

DIAGNOSIS AND PREDICTION OF HEART DISEASE WITH NEURAL NETWORK

Sushma C V[#1], Dr. Jyothi S Nayak
[1] *Department of Computer Science and engineering, BMSCE, Bangalore, India*

Abstract— People living in urban cities are highly prone to risk of getting chronic diseases. An early visualization of cardiovascular disease helps in having choices to way of life changes in higher risk of patients and thus decrease their confusions. The healthcare industry produces enormous amount of data daily in terms of vital parameters of patients. These data can be used for early prediction of the disease. The patient's demographics along with the symptoms can make the prediction more reliable. The ensemble of predictions based on both vital parameters and symptoms is proposed for the patients whose clinical data is available. The prediction tries to use only the symptoms as the input for those living in rural area as they cannot afford the cost for diagnosis To use the symptoms in the prediction model, the natural language processing where word embedding and also network modelling are combined which forms informative representation with the help of GloVe algorithm.

Keywords—preprocessing, Heart disease, Deep Neural Network, classification, Recurrent Neural Network, Global Vector, Logistic regression.

I. INTRODUCTION

An expanding volume by means of social insurance information which are contained in Electronic Health Records made many to think about planning computerized clinical help and malady identification frameworks dependent on patient history and hazard factors [15]. Various past examinations have endeavored to utilize persistent research facility tests judgments and drugs as methods for anticipating ailment beginning. Such models have likewise been utilized to recognize possibly obscure hazard factors regularly while all the while enhancing sensitivity and specificity of detection.

Various ongoing examinations have been fruitful in foreseeing sickness by means of different strategies including support vector machines,[5] random forests, neural networks[11], logistic regression, and time series model techniques. For both data representation and diagnosis in medicine deep learning methods are successful particularly.

Also, this venture adds to the momentum writing by giving an extensive study of different machine learning calculations on disease forecast undertakings. The medicinal services ventures gather hidden measures of information that contain some shrouded data, which is valuable for making successful choices. Data mining technique is used for providing appropriate results and making effective decisions. utilizing neural system for foreseeing the hazard dimension of coronary illness. The system uses medical parameters such as age, cholesterol, sex, blood pressure, for predicting the disease [12]. The number of studies are being successful in predicting disease. A powerful coronary illness forecast framework is created with methodologies namely, random forests, support vector machines, logistic regression, neural networks and time series models which are Long Short Term Memory and Recurrent Neural Network (RNN).

A. PROBLEM STATEMENT

The healthcare or social insurance industry creates a lot of medicinal services information day by day which to extracts information for predicting disease. The extraction of information from that of data is the major challenge. The scope of the paper is to detect disease at an early stage.

II. RELATED WORK

[1] The research paper proposes an approach that makes use of disease prediction model. The number of approaches is applied for predicting future disease with the help of neural network. This research paper proposes an approach where the database in order to predict heart disease and the dataset has the total of 300 records where it is parted as a set of training data and a set of testing data. The set of training data has 40% and the set of testing data has 60%.In this paper Weka tool is used and also Multilayer Perceptron Neural Network (MLPNN) along with the backpropagation is employed. Then classification of data is done and the confusion matrix is prepared which has information on real and predicted classification. Cleaning and filtering are applied to the dataset to remove unwanted data and filter the irrelevant data from the database. This makes use of patient's historical data and it determines the exact pattern and their relationship with heart disease.

Results: Multilayer perceptron neuron network model gives the better results. Even without retraining the system it works well.

[2] This research paper has proposed various techniques. The study made through cardiology which has the dataset of 251 patients. In preprocessing the dataset contains missing values. The dataset which are preprocessed has 217 cases and 76 features. The classifiers such as KNN and SVM are employed. Random forests utilize different tree classifier which is worked from an example of predictors to lessen change to single arrangement trees. So in order to find minimal features which are used for this classification feature selection are employed.

Results: The methods proposed in this system are complex and it involves time consumption and the operational time that it takes is little high.

[3] In this paper the word embedding techniques are employed which uses natural language processing. The approach used here is to take the patient's record and encoding of that record is represented and try to predict the disease in the interval of its time of [t+1,t+n] based on patient's next arrival. This paper involves convolutional neural network, Recurrent neural network. The embedded representations adopt Glove algorithm. The vectors are obtained from the embedding layer and are fed to recurrent neural networks. Initially disease is predicted and data preprocessing is applied to various set of diseases like heart failure, diabetes mellitus, chronic kidney disease. The performance is evaluated using logistic regression, random forest and XGBoost. Results: XGBoost performs better than LSTM.

A recurrent neural network [8] is a deep learning concept that is a class of deep neural network. It sees application in visual imagery analysis .A variation of multilayer perceptron is used, designed with the intention of minimizing preprocessing. The convolutional neural networks are supervised and they use labels as learning signals. Recurrent Neural Network have working like that of a feedforward network but there is temporal dependency between inputs as it is a sequence unlike the traditional feedforward neural network. Here output not only depends on previous input but also on the whole history of input. But the problem is that the gradients vanish because of which it is not practical to look back for long sequences. This results in the introduction of LSTM (Long Short-Term Memory) which incorporates the forget gate which happens to be the recurrent gate. This prevents vanishing and explosion of back propagated errors. Thus it is made practical to work on longer sequences that are stacked together for higher level of information capture.

III. PROPOSED SYSTEM

A. *Pre-Processing:*

Preprocessing is an initial and vital process. It is the data mining strategy [1] that changes raw data into a reasonable configuration. Raw data (real world information) is constantly inadequate and that information can't be sent through a model. That would cause certain blunders. That is the reason we have to preprocess information before sending through a model. The first steps involved in preprocessing is import the libraries and second step is to read the data and third step is to check for missing values and fourth step is to check for categorical data and fifth step is to standardize the data and then PCA transformation and at last data splitting. Pandas, time and Numpy are used as main libraries. Pandas for manipulation of data and analysis of data. Numpy for scientific computing with the help of python. Matplotlib and seaborn for visualization and for preprocessing technique scikit-learn is used. Then the data is read from the dataset and checked for missing values. PCA (Principal Component Analysis) fundamentally utilizing to decrease the span of the element space while holding however much of the data as could reasonably be expected.

B. *Word Embedding technique*

Global Vectors[1] called as GloVe algorithm is applied for the dataset .It is a model which describes distributed word representation .The model is unsupervised learning where in obtains vector representation for words. There is mapping of words with real values numerical. Then on word-word cooccurrences statistics from the corpus training is done.

The GloVe display is prepared on the non-zero sections of a worldwide word-word co-event matrix, which organizes how every now and again words co-happen with each other in a given corpus. Populating this framework requires a solitary go through the whole corpus to gather the measurements.

The first step is to gather word co-occurrence insights in a type of word co-occurrence grid X. Every component Xij of such network speaks to how regularly word i shows up in setting of word j. The soft constraints are defined for word pair and then a cost function. As a rule, a co-occurrence matrix will have explicit entities in lines (ER) and columns (EC). The motivation behind this matrix is to display the occasions every ER shows up in a similar setting as every EC.

C. *Recurrent Neural Network*

In Recurrent Neural Network, the output has been fed into input over present or current step. So, in the context to predict the next word in a sentence, it is required to have the previous result or word .So there is a need to remember the words of previous state where there is a dependency. So Recurrent neural network have a hidden layer .The hidden layer has hidden state associated with it which is used to remember some sequence of data.

RNN [19] constitutes the memory which recalls data about what has determined. And it utilizes indistinguishable parameters, from it plays out a similar assignment on every one of the sources of info or shrouded layers to create the output to each of the input. RNN changes over the autonomous activations into dependent initiations by giving out similar weights and bias to every one of the layers, along these lines lessening the multifaceted nature of expanding parameters and retaining each past yields to each previous as contribution to following hidden layers.

The formulae which calculates the current or present state is

$Ht=f(Ht-1,Xt)$

Where Ht represents the current state Ht-1 represents previous state

Xt represents the input state

The formula for activation function is

$Ht=\tanh(WhhHt-1+WxhXt)$

Where

Whh represents recurrent neuron weight

Wxh represents at input neuron weight

The formula for calculating output $Yt=WhyHt$

Where

Yt is the output

Why represents output layer weight

Training

i) An input is provided to the network ie., a solitary time venture of the information is given to the system.

ii) At that point figure its present state utilizing current info and past state. And the present state Ht progresses toward becoming Ht-1 whenever step.

iii) Some can go the same number of time ventures as per the issue and combine all of data which belongs to past state.

iv) The last state is asserted as an output when the steps for time are finished.

v) An error is created and output is contrasted with real output.

vi)Then an error is back-proliferated to system in order of refreshing the loads and consequently the system or neural network is trained.

A Recurrent Neural Network recollects every single data through time. It is valuable in time arrangement forecast simply because of the component to recall past contributions too. This is Long Short Term Memory.

Recurrent neural network are utilized with layers called convolution to broaden the powerful pixel neighborhood. In the base model where ReLU activation which is called as Rectified Linear unit is applied .The activation function employed is the rectified linear unit in deep learning models. For any positive value x, this function returns the value back otherwise the function returns 0.Therefore it is represented as f(x)=max (0,x).This function allows the model to have non linearity. Here non linearity is that the slope is not constant.

Many models has bias for each node. This is used during model training which is a constant number. For example, consider a node with input X and bias .If the bias takes a value of 5 then the node output is f(5+A).Here if X is greater than -5 then node output is 5+A and has slope of 1,if X is less than -5 then output is 0 and slope is also 0. So the bias term enables us to move where the slant changes. Up until now, regardless it shows up we can have just two distinct inclines. Be that as it may, genuine models have numerous nodes. Every node (even inside a solitary layer) can have an alternate an incentive for its inclination, so every node has the ability to change slope at various inputs. When we include the subsequent functions back up, we get a consolidated function that changes slopes in numerous spots. These models have the adaptability to create non-linear functions and record for communications well. As we include more nodes in each layer the model gets considerably more noteworthy capacity to speak to these connections and non-linearity. For the dense layer of 100 units there is sigmoid activation. Sigmoid activation exists in the range of 0 to 1

.The probability lies between 0 and 1 in order to predict. Here, slope of sigmoid curve can be found which is differentiable. The sigmoid function looks like a S-shape. The activation functions are generally used in neural network. Basically it determines the output of neural network.

In the following preprocessing and organization of data is done. The records of each of the past patients are well organized and it is fixed for the final or target year.

1) Framing of the Features: The diagnoses systems marker cross section is utilized to screen which assurance occurs with which strategy.

2) Rundown of the Features complexity in the past history of the Patient: Undefined approach from with heart afflictions is used: There are four time hinders for each therapeutic factor with each and every relating record consolidated multiple, four or five years already goal year, and the next time normal midpoints of all the earlier records are there.

3) Part the Data into a Training Set and a Test Set Randomly: In basic Artificial Intelligence is being administered, the data are part into a training and a test set. From the perspective, all of the patients' information are taken from the similar distribution, and are not distinguished with those patients whose records are there apriori with those of later ones. Training set of data are more as compared to testing data. The model is trained on different set of values to bet better accuracy.

D. Random Forest

The classification and regression problems are solved with random forest algorithm. Random forest algorithm comes under the classification of supervised algorithm.

As the name propose, this calculation makes the forest with various trees. The model is robust if there are more number of trees and it gives more accuracy. Random forest classifier deals with the missing qualities. It is a machine learning algorithm. Here there will be troupe of decision trees and rather than looking for the most significant component while part a node, it scans for the element with different set of features. Thusly, in Random Forest, just an irregular feature of the subset is thought about by finding for part a node. It is possible to even make trees increasingly arbitrary, by moreover utilizing irregular thresholds. It is simple to predict the feature by random forest and it is one of the nature of calculation.

E. k- Nearest Neighbor

KNN is used for prediction which rely on the mean or the median of K instances for classification using regression. KNN stores the training dataset which it utilizes as its portrayal. This predicts just in time where it calculates the resemblance between each of the training data and the sample of input. This is one of the supervised algorithm.

F. Classification:

The final process is the classification which classifies the result. Training set is the one which user knows. Test set is the one which user want the method of classification. Test informational collection utilizes the preparation informational index utilizing which an outcome will be yield that tells whether the given dataset has a danger of coronary illness.

IV. METHODOLOGY

A. Using Logistic regression

Step 1: Data collection and dataset preparation

Various hospitals has huge information about patient health. So, there is a collection of the data from all sources which may include different hospitals, receipts, or from repository and after that preprocessing is done in order to remove missing data or eliminate redundant data and eliminate all extra information. The dataset is collected from the hospitals which has 15 attributes in it and 4500 records. The goal is to predict the future risk of getting heart disease. Each of the attributes includes patients behavioral, whether the person smokes daily or not, medical such as blood pressure, hypertension, cholesterol level, medical such as body mass index, heart rate. Initially preprocessing is done to remove null values, missing values, redundant values on the set of attributes.

Step 2: Develop a model for logistic regression approach for Disease Prediction.

Here, logistic regression approach is employed for prediction of heart disease from its attributes and probabilistic model is developed based on its attributes. The data present in database contains huge number of records with attributes

It is supervised leaning. Here the number of attributes of the patient maps to the output of the result whether the patient has the disease or not.

Step 3: Training and testing on datasets

The model is trained on the dataset of values and does the prediction accurately. For logistic regression there it takes the training set as 60% whereas testing set has 40%.The model is trained by varying these values.

The model has some attributes with p value which is greater than alpha., which can be taken as 5%. This shows there is probability of getting heart disease with statistical lowest relation. Backward propagation is applied to eliminate the value of p greater than alpha. The model is run on all the values to fit. This will give the result of p value less than alpha.

B. Using Recurrent Neural Network

Step 1: Data collection and dataset preparation

The text dataset is collected which has information of number of diseases, count of the number of diseases and symptoms of corresponding disease.

Step 2: Develop a model and deep learning approach for Disease Prediction.

Here, the deep learning approach is employed for prediction of heart disease from its symptoms and probabilistic model is developed using logistic regression based on its attributes respectively.

The data present in database contains huge number of records with attributes and other dataset has symptoms that can be used for predicting disease

For predicting disease with symptoms the raw text is the dataset used. The prediction is made based on word. That is tokenize of words are done by splitting the words. The infrequent words are removed. The vocabulary size is limited

Step 3: Implementation of Recurrent Neural Network

For predicting heart disease with symptoms, there are 1866 symptoms. For the first array, a vector representation of these symptoms is given as input. The output is a vector.

There are 1866 symptoms and 34 diseases taken in the dataset. In general, Recurrent Neural Network is a procedural arrangement and it will process the sequence ,For example stock price prediction(daily occurrence) thereby retaining the memory by considering the previous result which had occurred in the sequence. The output at the present time step turns into the contribution to whenever step or the next. So the model remembers the previous output as well. As human thinks likewise RNN mimics the way. There is a dependency of long term.RNN provides with the gradient to make sure that the signal is not lost while processing. It will maintain the current state.

The dataset has sequence of words that is given as input to the model. Here the input is the symptoms. The model will be trained to predict the next outcome(diseases).These each symptoms are mapped to integers. Using embedding matrix these integers are mapped to vectors and then passed to the Recurrent Neural Network. So here the symptoms are passed as input parameters and then prediction is done. Then update the input parameter and other prediction is done and it proceeds likewise.

The initial step is to convert the list of words to list of integers. Then features and labels are from those of sequences .Then build the RNN model using embedding and Dense layer. Pre trained embeddings are loaded .Then the model will be trained to predict the next disease and make the prediction.

So here in order to convert the words or individual text to an integers Tokenizer is used with Keras. So this converts each of the words to integer sequences. This is done using sequence= tokenizer.texttosequence(text). Then idx_word indicates what these integers represents ie., mapping of index to word. So next step is to train the model. Here 50 are used as feature and 51 as the label, this gives the performance for the network because it is proportional to training the amount of the data. The embedding is used to map input word to a 300 dimensional vector. The embedding takes weights as parameter. There is masking layer involved to mask words (that do not have data

pretrained).The 300 dimension embeddings are from Global Vectors algorithm.

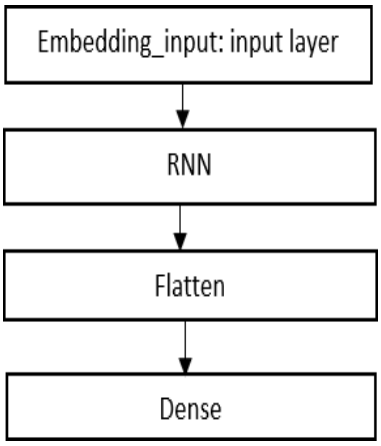

Fig1: Prediction of the heart disease with GloVe

The embeddings of pretrainable contains over 400,000 words. By loading these embeddings, embedding matrix can be formed.

Flatten is a layer of the neural network that changes the output of RNN to a one dimensional array (basically flattening the dimension).Dense is a regular densely connected neural network that translates X inputs into Y outputs.

Then the model is optimized to get better accuracy. Dropout allows to basically turn off certain nodes in the neural network. Thus allowing the model to readjust. The overall goal is to prevent the model from becoming overconfident. Here, overconfident means, it assumes a lot thus resulting in bad predictions and starts looking for the same patterns in the data (over trained model).

The vectors are the input to recurrent neural network. So mapping is created with those of indices and words. The input is a sequence of words and each word is a vector of defined vocabulary size. So there are 40000 rows with 300 columns. This has been trained on 50 epochs .The embedding layer is the layer which maps each of the input text to a 300 dimensional vector. The pretrained weights are employed for embedding layer, and then the masking is applied for those words not used in training. Then the cells with RNN with dropouts are used to prevent the model from becoming overfitting.

Keras is the python library used for deep learning. It performs well for numerical calculation and for creating neural network model. The iterations are continued with the dataset which are limited and fixed. The model is fitted on the data. The model is evaluated on the training dataset. Then it generates prediction for each of the input.

Step 4: Deployment of the model

The model will be conveyed in a genuine situation made by the human specialists and will be utilized for improving further.

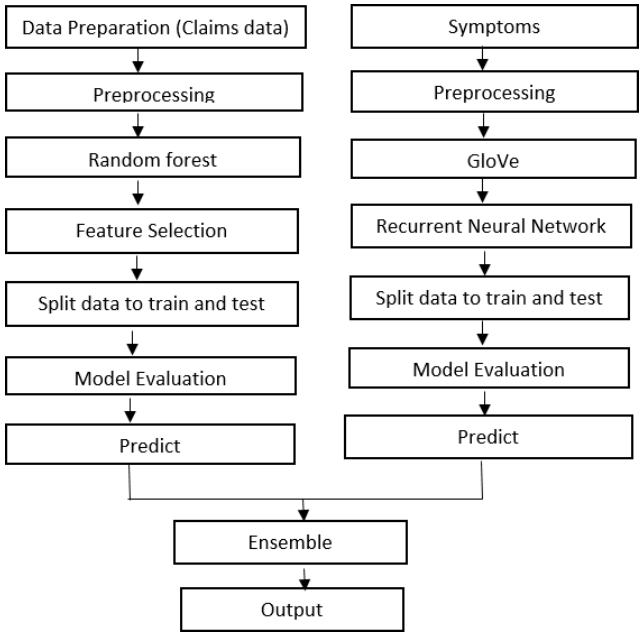

Fig 2: The overall project architecture

RESULT ANALYSIS

The main aim of our research is to predict heart disease, so here we are obtaining approximately 1800 users data from dataset for predicting using regression technique.

We performed preprocessing by removing null values, missing values, redundant values in the data.

Later the regression is performed and true positive, true negative, false positive and false negative is calculated for classification. The true positive rate or sensitivity and true negative rate or specificity is calculated for evaluating model.

Then for evaluation matrix we used back propagation and confusion matrix concept and found the score and roc curve of the algorithm as shown in fig 3, fig4. The roc curve for the algorithm is 0.73. Fig 4, shows the confusion matrix score for TP, TN, FP, FN as seen below. The rate of True positive is 43.47%, True negative is 99%, False Positive is 80% and false negative is 88 %.

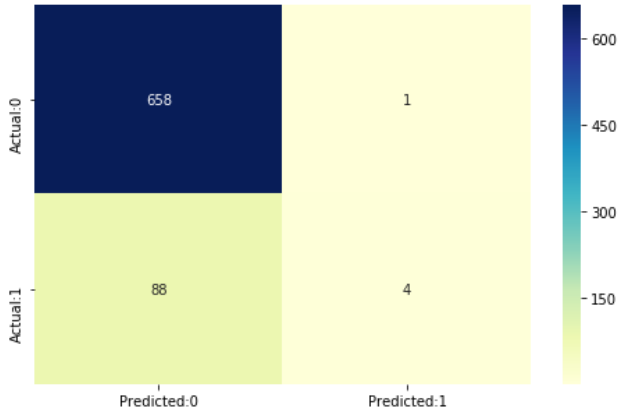

Fig3:Confusion matrix

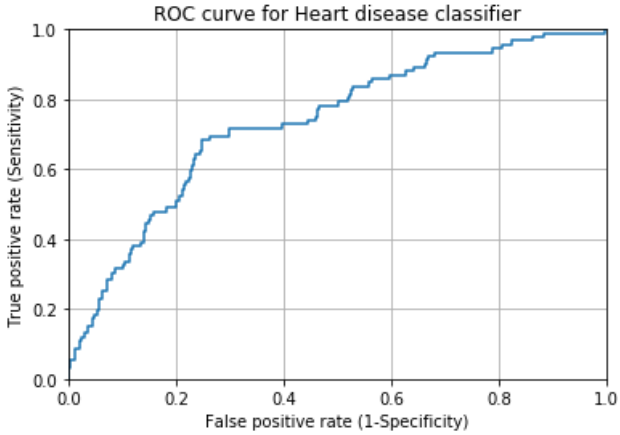

Fig4: ROC curve

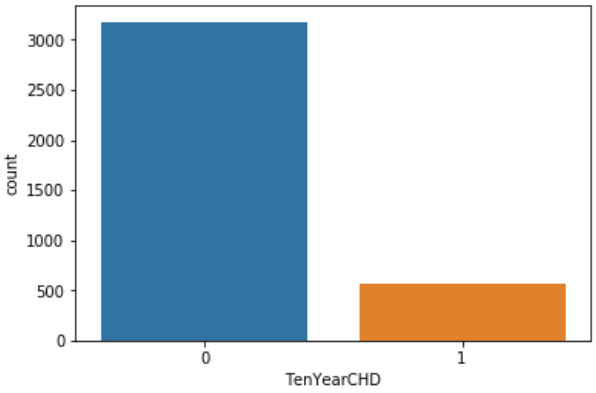

Fig5: This shows the number of people who are not at risk of getting heart disease(indicated by blue in color) and people who are at high risk of getting disease (indicated by red color)

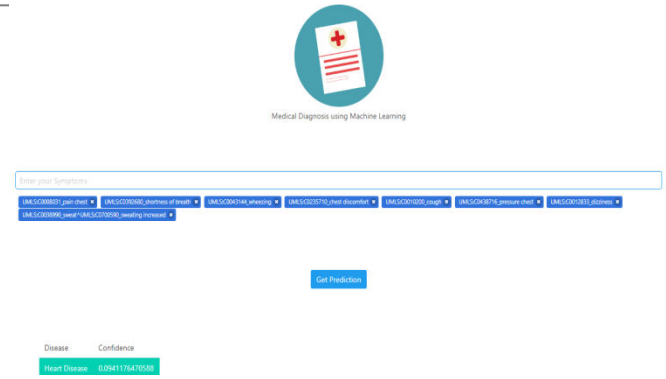

Fig6: Prediction of heart disease with Recurrent neural network by integrating it into Django framework

I. CONCLUSION

The goal of the project is to develop a system that predicts heart disease at the early stage. The proposed prediction technique is inputted with the dataset of the patients. At the urban cities, people are diagnosed with the symptoms as well as with claims data by means of logistic regression but at rural region, as they cannot afford money are diagnosed with symptoms by means of embedding technique and are predicted.

II. REFERENCES

[1]Christensen, T., Frandsen, A., Glazier, S., Humpherys, J. and Kartchner, D., 2018, June. Machine Learning Ways for Malady Prediction with Claims Knowledge. In 2018 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 467-4674). IEEE.

[2]N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[3]R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis," arXiv preprint arXiv:1608.02158, 2016.

[4]Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in 2017 IEEE International Conference on Data Mining (ICDM), Nov 2017, pp. 787–792.

[5]Vembandasamy et al., "Heart Diseases Detection Using Naive Bayes Algorithm", vol.2, pp. 441-444, 2015

[6]Otoom A.F., Abdallah E.E., Kilani Y., Kefaye A., "Effective Diagnosis and Monitoring of Heart Disease", International Journal of Software Engineering and Its Applications, vol.9, pp.143-156, 2015.

[7]Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. IEEE T. Pattern Anal. Mach. Intell. 35, 1798–1828 (2013)

[8]Kumari, V.A. an d R. Chitra, "Classification of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications, vol.3, pp. 1797- 1801, 2013.

[9]MIN CHEN, YIXUE HAO, KAI HWANG,LU WANG and LIN WANG "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities" " IEEE Transactions on Healthcare big data, vol. 5, no. 4, pp. 1114– 1120, April 2017

[10]Lipton, Zachary C., David C. Kale, Charles Elkan, and Randall Wetzel. "Learning to diagnose with LSTM repeated neural networks." arXiv preprint arXiv:1511.03677 (2015).

[11]Huang, De-shuang. "Radial basis probabilistic neural networks: Model and application." International Journal of Pattern Recognition and Artificial Intelligence 13, no. 07(1999): 1083-1101.

[12]Choi, E., Bahadori, M. T., Schuetz, A., Stewart,W. F., & Sun, J. (2016, December). Doctor ai: Predicting clinical events via perrenial neural network. In Machine Learning for Healthcare Conference (pp. 301-318).

[13]Esteban, Cristóbal, et al. "Predicting clinical events by combining static and dynamic info victimization perrenial neural networks." 2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2016.

[14]QURESHI, KHAVER N., et al. "Neural network analysis of clinic pathological and molecular markers in bladder cancer." The Journal of urology 163.2 (2000): 630-633.

[15]Friedman, N., Linial, M., Nachman, I. and Pe'er, D., 2000. Using Bayesian networks to analyze expression data. Journal of computational biology, 7(3-4), pp.601-620.

[16]Xing, Yanwei, Jie Wang, and Zhihong Zhao. "Combination data processing strategies with new medical information to predicting outcome of coronary cardiopathy." In 2007 International Conference on Convergence Information Technology (ICCIT 2007), pp. 868-872. IEEE, 2007.

[17]Snow, Peter B., Deborah S. Smith, and William J. Catalona. "Artificial neural networks within the designation and prognosis of prostate cancer: a pilot study." The Journal of urology 152, no. 5 Part 2 (1994): 1923-1926.

[18]Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. "Using recurrent neural network models for early detection of coronary failure onset." Journal of the American Medical Informatics Association 24, no. 2 (2016): 361-370.

[19]Gokul, S., M. Sivachitra, and S. Vijayachitra. "Parkinson'smaladyprediction victimization machine learning approaches." In 2013 Fifth International Conference on Advanced Computing (ICoAC), pp. 246-252. IEEE, 2013.

[20]Meccoci,Patrizia,,Enzo Grossi, Massimo Buscema, Marco Intraligi, Rita Savarè, Patrizia Rinaldi, Antonio Cherubini, and Umberto Senin. "Use of artificial networks in clinical trials: a pilot study to predict responsiveness to donepezil in Alzheimer's disease." Journal of the American medical Specialty society fifty,no.11(2002):1857- 1860

[21] Jin, B., Che, C., Liu, Z., Zhang, S., Yin, X. and Wei, X., 2018. Predicting the risk of heart failure with sequential data

Modeling,IEEEAccess,6,pp.9256-92