Disease Prediction by Using Deep Learning Based on Patient Treatment History

Kadam Vinay R, K.L.S.Soujanya, Preety Singh

Abstract— Now a day's medical and healthcare industries has big business. Healthcare industry produces large amount of data in daily rou-tines. That big amount of data is used for the future disease prediction. Prediction is does on the patient previous history and health related information. In every hospital or clinics there are large amount of patient history or patient related all information is available. But there is the major challenge is that how to extract that information from that large data records. For doing the prediction of diseases using that patient treatment history by applying the machine learning and the data mining techniques is the continuous struggle for the past few years in medical or healthcare industry. In previous years many papers are published by using the data mining techniques and machine learning technique for the disease prediction, also for the progression and reoccurrence of those diseases. In this paper we build the new model for the diseases prediction. In that we use the deep learning concept artificial neural network (ANN) for predicting the diseases. In this paper we use the probabilistic modelling and deep learning concept for prediction. For that we collect the three diseases heart, kidney, and diabetic's dataset. For those diseases we build the one proper dataset. That dataset are split into the training and testing dataset. For training the dataset we use the scholastic gradient decent algorithm. For this project we collect patient related information. We collect the datasets from the UCI Repository, Pima datasets, and Kaggle datasets. For that collected dataset we apply the pre-processing and remove the unnecessary data and extract the important features from that data. On the new generated data we apply the probabilistic model and deep learning technique for doing the diseases prediction. Then by using ANN method we do the prediction and produce the confusion matrix. Then trained and tested model will be deployed in a real-life scenario for diseases prediction. We got the prediction accuracy 95%, 98%, 72% for heart, kidney, and diabetes disease respectively which is far higher when compared to the existing methods.

Keywords: Health-care, Deep Learning, Artificial Neural Network, Disease Prediction, Health Data.

1. INTRODUCTION

"Prevention is better than cure", is universal truth. In the human life Health is the most important factor. So now a days there is need to do the prediction of diseases. Many researchers have used data mining and machine learning techniques for predicting the diseases based on the medical data or pathological data [1]. These approaches are used for doing the prediction of diseases and reoccurrence of those diseases. Also, some another approaches used to predict the diseases and control the diseases. This approach also controls the progression of particular diseases.

The recent success of deep learning in different areas of machine learning has driven a shift towards model of machine learning that can learn hierarchical and rich representations of raw data with the pre-processing and produce more accurate result [1]. Diseases related numbers of papers have been published on several data mining and machine learning techniques. In that various data mining and machine learning techniques are used like Naive Bayes algorithm, neural network, decision Tree algorithm, K-nearest neighbors algorithm bagging algorithm, so on [2].Also Heart related numbers of papers are published using data mining and machine learning techniques [5]. Those techniques such as support vector machine, kernel density, automatically defined of heart groups for showing different levels of accuracy in diseases prediction [20]. Generally Waikato Environment for Knowledge Analysis (WEKA) tool is used in this type of researches.

Previously existing systems do the prediction on diseases but cannot predict the subtypes of diseases. Those systems also can't find the diseases which are caused by occurrences of any previously diseases. Those systems fail to predict possible conditions of people. Previous system can handle only structured data but not an unstructured data. In current past, countless disease estimate classifications have been advanced. The standing organizations arrange a machine learning algorithms which can predict exact diseases.

In the proposed system we use the artificial neural network (ANN) and stochastic gradient algorithm for learning and doing the effective prediction of diseases. This system handles the both structured data and unstructured data with the help of preprocessing. For that we collect the patient previous history like patient diseases details. Also we can collect the various data sets from UCI Repository, Kaggle, dataset data.gov, and Pima datasets. By using that collected data we prepared new dataset. In our project we contain the three diseases like heart, kidney, and diabetics. We collect the all diseases related data. We combined that dataset into one dataset. In that we first take the common attributes from that datasets and also take the some another important attributes related to that three diseases. For that combined dataset we find the all missing values by using the decision tree liner regression method and generate the missing values. Finally we build the proper dataset. By using that dataset we build the model with artificial neural network (ANN) for predicting the decease and do the predictions using proposed model with better accuracy.

Revised Version Manuscript Received on March 10, 2019.

R. Kadam Vinay, M .Tech Student, Department of CSE,CMR College of Engineering and Technology, Hyderabad, India. (E-Mail: kadamvinay55@gmail.com)

K.L.S. Soujanya, Professor, Department of CSE,CMR College of Engineering and Technology, Hyderabad, India. (E-Mail: klssoujanya@cmrcet.org)

Preety Singh, Assistant Professor, Department of CSE,CMR College of Engineering and Technology, Hyderabad, India. (E-mail: preeti17singh@gmail.com)

1.1 Importance of diseases and its types

Now a days in world diseases percentage is increased because of the change in the weather, change in human habits' and many more reasons. There is the need for preventing the diseases and also predicting the diseases. In the world every human have some small or big health related problem. So there is needed to take care from those problems. In the world many more diseases are present that cause the human life.

Table1. Diseases with attributes

Sr. No	Diseases	Attributes
1.	Heart	Age, Sex, Blood Pressure, Blood Sugar, chest pain, Cholesterol, restecorg (resting electrocardiographic), thalach (maximum heart rate), exang (exercise included by angina)
2.	Kidney	Age, Sex, Blood Pressure, Blood Sugar, sg (specific gravity), al(albumin), rbc (red blood cells), sc(Serum)
3.	Diabetes	Age, Sex, Blood Pressure, Blood Sugar, BMI (Body Mass Index)

In that project we discuss about the few diseases. For that we do the early prediction for avoiding and take caring of that disease. In this project we take the three types of diseases like heart, kidney, and diabetes diseases shown in the Table 1. In that we show the three diseases with its attributes like age, sex, blood pressure, blood sugar, Cholesterol, restecorg, thalach , exang, sg, al, rbc, sc, BMI for the heart, kidney and diabetes.

Now a day's those three diseases are very common in the human life. Many people are dies because the heart and kidney diseases because those diseases treatment are very costly. For that we can build the new model based on the artificial neural network for predicting those diseases early. By predicting diseases early we can gives the good treatment to that particular patient and save the life of that person in less cost.

1.2. Machine learning

Ml algorithm can be helps to provides vital statistic, advanced analytics, real time data in the terms of patient's diseases, family history of patients blood pressure, lab tests results, clinical data and more to the doctors. In a healthcare industry ml is the fastest growing trends. Ml technologies can also help to medical expert's people to analyze the data for improving diagnoses and treatments of diseases. Ml can be used in healthcare and life science for disease identification and risk satisfaction about the particular patient diseases. It also used for diseases identification with speed and accuracy in healthcare organizations. ML also used for diagnosis in medical imaging for showing the more complete image of diseases or illness. It is used in medical drug discovery and robotic surgical tools.

1.3. Deep Learning

In healthcare organization deep learning is used for uncover the hidden opportunities and patterns in clinical data helps to doctors for treat the patients well. Deep learning collects the huge amount of data, including patient related records data, medical reports of patients and on that data applies neural network techniques to provide better outcomes. Deep learning solves the problems which are not solved by the machine learning. Deep learning uses the various neural network methods and provides the better results. In healthcare deep learning provides the analysis of any diseases accurately to doctors and helps to doctors to treats particular patient and give better medical decisions. Deep learning techniques analyze the patient's medical history or health data and provide the best treatment for that patient. Also deep learning methods are used for Alzheimer diseases at an early stage of that particular person.

1.4 Artificial intelligence

In medical field AI is the great idea that can improve the communication between patients –doctors and healthcare professionals in health organization. By using the AI stores the large amount of data and processes that data on standard manner. With the help of the AI avoids the human errors comes in past. Some diseases require immediate actions over it otherwise they will become more rigid; by using AI those diseases are easy to handled. By using artificial neural network such problems are solved fast and generate the accurate result. With the help of AI techniques patients can get the doctor assistance without visiting to hospitals and clinics which was reduce the cost of patient treatments. By using AI early detection of the various diseases can reduce the grimness of diseases.

2. LITERATURE SURVEY

Min Chen et al., [1] developed techniques for predicting the diseases with the help of machine learning. They can propose new techniques based on the machine learning concept with the help of convoluational neural network. They proposed new method as multimodal disease risk prediction (MDRP) for predicting the chronic diseases. By using MDRP methods chronic diseases are effectively predicted .with the help of structured and unstructured data they can predict the diseases. They use machine learning and deep learning algorithms for prediction. In that machine learning algorithm such as k-nearest neighbor, naive Bayesian and decision algorithms and deep learning algorithm convoluational neural network used to predicting the diseases risk. MDRP algorithm process the datasets into two parts as training set and testing set which is train and test the data respectively for better prediction of diseases with good accuracy. By using this technique predict the whether the patient have a chronic diseases or not. The predicting accuracy of proposed algorithm is 94.8% with the high speed of predicting the diseases. But with the help of convoluational neural network it is difficult to determine window size of data and it can't handle the sequential data.

R.Tamilarasi et al., [2] proposed a system for predicting hearth diseases with the help of data mining techniques. In medical science large amount data is generated from patient clinical reports other patient symptoms. Data mining is used to handle that large amount of data with the help of classification and clustering. They studied different data mining techniques that can useful to predict the heart diseases. Such data mining classifiers technique are used for effective and efficient heart diseases diagnosis. In that they use various attributes and decision tress method for predicting diseases. Data mining techniques are used to analyze the data from different dimension and identify their relationships. For predicting the diseases they uses data mining algorithms like decision tree algorithm, naive bayes algorithm, neural network algorithm, k- nearest neighbor algorithm with the classification of diseases. This data mining techniques helps to healthcare professional for diagnosis of heart disease with better accuracy. Proposed system accuracy is 85%. But some disadvantages of data mining techniques like they are lazy.

Darcy A. Davis et al., [3] proposed the method for predicting the diseases which is based on patient medical history. They propose a CARE, collaborative Assessment and Recommendation Engine which depends on the medical history of patient. They use the IDC-9-CM codes to predict the diseases risks. This method is used for predicting the chronic diseases. In that they also describes a Iterative version of CARE, as ICARE which is incorporates ensembles concepts, but those approach did not have positive capacity of prediction. CARE system can do the prediction based on the vector similarity, inverse frequency and clustering with the medical data of patients. In that IDC-9-CM is the 3-digit code, which represents the small group of similar or related diseases of patients. CARE framework is used explore the border history of diseases suggestions related to previous unconsidered concerns about the prevention. But CARE system generates prediction on only feature visits of patients based on medical history.

Feixiang Huang et al., [4] developed a system for predicting the diseases by using data mining techniques with the healthcare information. For that they apply data mining process which predicts the hypertension of patient by using the patient medical records. In that 9862 sample cases are studied. This sample is extracted from the real word information system databases. That information system databases contain 309383 medical records is used to generate diseases prediction. For that prediction data mining techniques are used such as naive Bayesian and RJ-48 classifiers. In that WEKA data mining tool is used to generate those data mining techniques. Confusion matrix is used to represent the performance of naive Bayesian algorithm. In that they use a simple approach of considering the present or absents of diseases in medical history of patient. Accuracy of proposed system is 83.5%.

Abhishek Rairikar et al., [5] proposed prediction model for predicting the heart diseases with the data mining techniques. In that they uses a different more numbers of patients attributes, such as gender , blood pressure, cholesterol like other some attributes for predicting the heart diseases. Healthcare industries produce massive volume of data which is forms of numbers, text, images, and charts. Data mining provides the various classification methods like K-nearest neighbor, decision tree, CART, C4.5, J48 and so on. In this system three different data mining classification techniques such as K-nearest neighbor, decision tree and naive bayes are used to analyze the datasets. K-nearest neighbor classification and regression methods are used to pattern reorganization and decision tree are used to build the good decision. But the KNN algorithm is lazy algorithm, where the functions are only locally approximated and also in that need to determine values of parameters of previous neighbor.

Saurabh Pandey et al., [6] developed efficient way to predicting the diabetes of patient by using the bio medical signal data with the help of artificial intelligence techniques. This system gives brief overview of diagnosis of diabetes using patient medical bio signal data. In that they use the artificial intelligence approach like ANN, Fuzzy for fixing the wide variety of issues in different application of area. They propose the suitable approach for prediction of diseases based on the dietetics bio medical signal data. Workflow of the methodology is like feature selection as symptoms of diseases then building the datasets with data homoscedasticity after that training and testing of datasets are done by using AI techniques. For the simulation result they use the algorithm which is developed by using MATLAB for detection of diabetes. For that datasets are used with the number of input value which is selected by using regression analysis. In that they use the 768 input samples in diabetes datasets. After that gives the value of regression coefficient which shows the output dependency of every input sample that gives the prediction of diseases. For accurately representing the statistical properties of real time data which is does not possible to predict diseases.

Dr. B.Srinivasan et al., [7] studied the data mining techniques for efficiently predicting the diseases in healthcare sectors. They can introduce the various data mining techniques which are useful in medical fields for better decision making related to the diseases. In medical filed huge amount of data produced like the patient details, diagnosis history and varies medications, such data is used to predicting the diseases by using data mining approaches. They introduce the data mining knowledge discovery for converting the low level data to high level data knowledge. For that data cleaning, data integration data selection, data transformation pattern evaluation, knowledge representation such steps are required. In that they studied various data mining technique like as, Bayesian classifiers, decision tree, support vector machine and artificial neural networks for predicting diseases. They discussed about various diseases like Eye diseases, Cancer, Heart diseases, Diabetics, etc. Data mining based prediction systems reduce the cost and human effects but they are time consuming and lazy learning methods.

Parithosh Khubchandani et al., [8] proposed a system based on artificial intelligence and probabilistic model for medical prediction. Prediction is the important factor in the medical domain. In that they can use the artificial intelligence for decision making in medical filed to predicting the

diseases. This system can generate the important data for the evolution of diseases diagnosis. Therefore main advantages of artificial intelligence are it creates tools that should better work than human. In that they present the new approach suitable for medical prediction which is based on the probabilistic modeling. When the information is large and complex the system uses this approach. Knowledge based approach cannot handle the large or complex data, so probabilistic approach is used to medical prediction. The statistical approach associates a probability each output of medical data for that bayes theorem is used for prediction. By using this system physicians can focus on important activities of patients. Some result of technique take more time to evaluation and some computations are complex that is effects on the other factors of prediction.

Smita .T et al., [9] developed an efficient algorithm for predicting the diseases with the help of multidimensional data. Main objectives of this system create a easy, fast, effective approach for diseases prediction. They introduced new hybrid algorithm for diseases identification and prediction by using data mining techniques. New algorithm is diseases identification and prediction (DIP) it is combination of decision tree and association rule. This is used for doing prediction of some diseases in particular area. Also it is shows the relationships between the different parameters of diseases. For that they use data mining approach for extracting the information which is previously not known. It also used for analyzing the information for prediction. This research work is based on different data mining approach on the multidirectional data analysis. For that they uses the common data mining models for prediction , such as Association rule, decision tree, clustering, classification rule and various statistical data mining tools. For DIP decision tree and Association rule which construct the Apriori principle. Apply the statistical mining techniques in cluster analysis for extracting the data.DIP predicts the diseases only on multidimensional data. The result is represented in graphical format.

Anandanadarajah Nishanth et al., [10] proposed a new method for early detection of the chronic kidney diseases by identifying the important features from the datasets. Chronic kidney diseases (CKD) are the not know those medical testes of patients are take for the other purposes that is useful for the diseases. In that they use the kidney dataset for identifying and detecting the kidney diseases. In that dataset various attributes are preset like the blood sugar, blood pressure, specific gravity, Albumin. Serum, blood glucose and so on. In this paper they use the different techniques for the detection like CSP and LDA. CPA is the Common Spatial Pattern and the LDA is the Linear Discriminnant Analysis which is identifies the important attributes and detects the chronic kidney diseases. In this paper classification methods is used for the identifying the attributes of diseases. This analysis contains the albumin, haemoglobin, specific gravity, haemoglobin, with the serum like important features for early detection of kidney diseases. With the Linear Discriminnant Analysis get the 98% accuracy of chronic kidney diseases.

Fatma Taher et al., [11] proposed a system for detecting the lung cancer by sing artificial neural network and fuzzy clustering method. Lung cancer is the common cause of death of people among the world because its symptoms are appears at only advanced stage. There are many techniques such as x-ray, CT scan, MRI is available for diagnosis of lung cancer but they are very expensive and time consuming. These systems solve the problem effectively with the help of artificial network and fuzzy clustering. For that they use a segmentation process which detects the lung cancer early stage. In that two segmentation techniques are used that is Hopfield Neural Network (HNN) and Fuzzy C-Mean (FCM). Hopfield Neural Network (HNN) it is the one of the artificial neural network which is used for image segmentation. Those propose the segmentation process for both black and white and colour images. HNN can very sensitive and it can detect the overlapping classes of images. Fuzzy C-Mean (FCM) is for fuzzy identification and pattern recognition which is based on the distance criteria. This algorithm contains a predefined numbers of inputs and gives the clusters of outputs. For input they can takes the number of image dataset of diseases and applies the both algorithm and gives the prediction on that image data. But this system is gives the result only on image datasets which requires more numbers of image datasets.

From the above survey it is observed that Disease prediction was done with traditional models such as machine learning, Data Mining using various algorithms like logistic regression, decision tree and so on. Previously existing techniques do the prediction on diseases but cannot predict the subtypes of diseases. Those systems also can't find the diseases which are caused by occurrences of any previously diseases. That fails to do predictions of all possible conditions of patients. Existing systems can handles the only structured data but ca not handles the unstructured data. Existing systems predicts the particular diseases with the help of data mining techniques which are ambiguous [18]. In the existing systems the datasets size is .small, for the patients and diseases are present some specific conditions and the characteristics are selected from experience. So such a pre selected characteristics are not satisfy the changes in the diseases and its influence factors. Those systems have a lower accuracy of diseases predictions. By using Data mining techniques algorithms such as KNN they consumes more time of prediction because those algorithms are lazy.

3. PROPOSED SYSTEM

Disease prediction by using deep learning based on patient treatment history which aims to predict the patient diseases based on previous historical data of that particular patient. In the present work we use the deep learning technique as artificial neural network for doing the effective prediction of diseases. For training model we use the stochastic gradient descent algorithm. The data is collected from various sources like hospital, clinics UCI Repository. We collect the patient previous history like patient diseases details, and medical reports as datasets. Also collect the standard datasets from UCI Repository, Pima datasets, Kaggle datasets dataset.data.gov, etc.

In the present study we took three diseases the heart disease, kidney disease, and diabetes disease. The dataset is obtained by doing preprocessing of the raw data.

The pre processing was done using decisiontree linear regression for finding the missing values. The preprocessed dataset is used for giving input to the proposed model. The proposed model is based on the artificial neural network (ANN). For doing the training we use the stochastic gradient decent algorithm. After training and testing we produce confusion matrix for generating accuracy of model. Finally deployed that the model for diseases prediction in real life scenario.

3.1. Techniques for Diseases Prediction

3.1.1. Artificial Neural Network (ANN)

Artificial neural network is the deep learning network which is inspired by the biological neural network which acts as similar to the human brain. In the deep learning neural network is not only network but it having a many different frameworks of machine learning algorithm. With the artificial neural network deep learning process the complex data inputs and learn that data and gives the effective output. The ANN network is the collection of the different node called as the artificial neurons which is connected with the each other. In that each node is transmitting the data signals to each other, artificial neurons receive that data signals and process that data. The goal of the neural network is to solve the problems in same way of human brain. Artificial neural network is the most interesting branch of artificial intelligence [16]. Basically ANN is the mathematical algorithm which is generated by using the computer. Artificial neural network learn from the standard input data and capture the valuable data and produce the output result like the human brain. Accuracy of problem solving by using ANN is better than other techniques.

Artificial neural network has the number of layers which are connected to each other. In the ANN there are the three layers as input layers, one or more hidden layers and the finally output layer. Input layers take the no of inputs and pass to the next layer of ANN which is the hidden layer. Hidden layer process that data and pass to next hidden layer if there is any hidden layer and again process that data and pass to the output layer. Output layer gives the effective and accurate result. Figure 1 shows the structure of artificial neural network. ANN can use the various activation functions like sigmoid, relu, softmax and tanh.

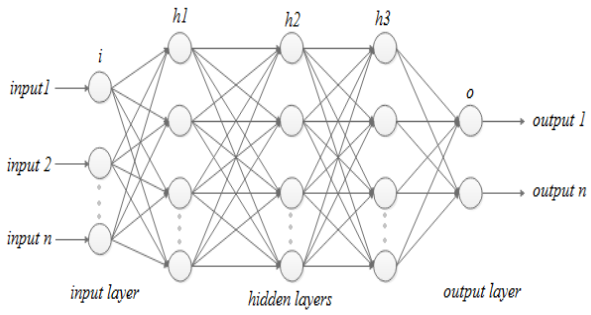

Figure 1 structure of artificial neural network.

In this project we build the artificial neural network by using the keras libraries in the python. In that we use the Sequential and Dense model for building the ANN network. For building the network first we initialize the ANN with the sequential model. After that by using the add method with Dense model we add the input layer and first hidden layer with the activation function. Then add the one more hidden layer with activation function in the network. Finally we add the output layer with the activation function. In that way we build the ANN model. Then we compile the network and fit the network with training algorithm. By using the training algorithm we calculate the loss and find the accuracy of model. Based on that model we do the prediction of diseases effectively. In the artificial neural network every linkage calculation are similar with the various activation functions. With the sigmoid, relu, Softmax activation function we can find the activation rate.

Equation of artificial neural network is,

$$y = f(w1x1 + w2x2 + w3x3 + w4x4 + \ldots\ldots + wnxn + b)$$

Here,

Y is the output.
w is the weight.
x is the input.
b is the bias.

3.1.2. Stochastic gradient descent algorithm

Stochastic gradient descent algorithm is a training algorithm which used for the train the model. In the deep learning and the machine learning it used for the train the huge amount of data sets. When training is carried with the gradient decent algorithm it is a very expensive process in the deep learning instead of that there is stochastic gradient decent algorithm is used. This algorithm is used for the find the local minima. By using this algorithm we update the weight after the execution of one row, then the next row for the all row in the dataset. By using the gradient decent algorithm we also train the deep learning model but they can't give faster result as compare to stochastic gradient descent algorithm. The Cost function for stochastic gradient descent algorithm as,

$$\text{Cost}\ (C, (y^i, y)) = \frac{1}{2}(y^i - y)^2$$

Here, y^i is the output value, y is the actual value, C is the cost.

The overall cost function of the stochastic gradient descent algorithm as,

$$\text{Train}\ (C) = \frac{1}{m}\sum_i^m (C, ((y^i - y)))$$

In this project we use the deep learning approach artificial neural network (ANN) for predicting the diseases. In that we build the model by using ANN, and for training purpose stochastic gradient descent algorithm is used. Some of the following steps are used for train the artificial neural network as,

Step1- Randomly initialize the weights to the network in input layer.

Step2- Give the first observation as first input from dataset, each feature in one input.

Step3- Do the forward propagation from the left to right. The neurons are activated in a way that the impact of each neuron activation limited by weight. And propagate the activation until the predicted output is not get.

Step4- After that we compare the predicted result to the actual result of dataset and measure the generated error.

Step5- Then do the back-propagation from the right to left and error is back-propagated. After that update the weights according to how many weights are responsible for generating the error. In that learning rate is decided how much we can update the weights.

Step6- Repeat the step1 to step5 and update the weights after the every observation.

Step7- when the whole training set data getting passed through in ANN then ANN makes the epoch and redoes more epochs.

3.1.3. Support vector machine (SVM)

Support vector machine (SVM) is the supervised learning model in the machine learning. Which is used as the learning algorithm for analyze the data which is used for classification and regression analysis. It is developed for binary type classification and latter used for the multiple classifications. It is the classifier defined for the hyper plane. Support vector machine algorithm effectively performs non-linear classification and also mapping the inputs in to the high dimensional feature space. Normally SVM constructs the set of hyper planes in high dimensional space which is used for classification, detection, and regression. In machine learning with the SVM training of data is done effectively for the classification problems.

In our project we use the support vector machine (SVM) for doing the prediction. We apply the SVM on the dataset and do the prediction. For applying the SVM we split the dataset in training and testing dataset. Then fit the SVM into training dataset. After fitting the data in SVM model do the prediction and evaluate the model. We check the prediction accuracy of the SVM model. But as compare to the ANN model we got the less accuracy for prediction

3.2. Methodology

Step 1- Collection of data and datasets preparation

This will involves the collection various medical related information is gathered from various sources. Data will be collected from the various sources like from UCI Repository. In that patient previous history, patient reports such information is collected. Also various diseases datasets are collected from UCI Repository, Kaggle datasets, Pima datasets, and datasets-data.gov. Then the pre-processing is applied on the collected dataset and extracts the important features and removes the unnecessary information from that datasets. By using the extracted information generate the new dataset which is used for prediction of diseases.

For our project we can prepare a datasets which contains three diseases as Heart, Kidney, and Diabetes disease. In that first we take conman features of all those diseases and also some other important features. By combining the all common and other features we make a dataset. But prepared dataset are having some missing values. There is need to find the missing values. For finding the missing values we use the decision tree linear regression. We build the model with decision tree linear regression with the known x and y values and also new x values by using that values we predict the new y values.

We split the dataset into training and testing datasets as x_train, y_train for training and x_test for testing and predict the y_pred as new values. By repeating that process we find the all missing values. We also tried with the multi regression and random forest linear regression methods but by using that methods prediction accuracy is very less as compare to decision tree linear regression method. By using decision tree method we get the greater than75% accuracy. So finally we used decision tree linear regression method and we build the proper dataset.

Step 2- Developing the new probabilistic modeling and deep learning approach (ANN) for diseases prediction

In this step develop the new probabilistic model with the help of deep learning approach that is artificial neural network for predicting the diseases. Artificial neural network is the deep learning techniques which is work as similar to human brain. With the ANN we build the new model for predicting the diseases. ANN runs effectively on prepared dataset. By applying the pre-processing we divide the dataset into the training set and the testing set. In that we divide dataset in 80-20% as for training and testing. On that divided training and testing datasets we apply the model and train by using the learning algorithm.

Step 3-training and experimentation on datasets

Developed diseases prediction model will be trained by using the stochastic gradient decent algorithm. By using stochastic gradient decent algorithm our diseases prediction model will be effectively trained on dataset. On that datasets ANN model do the accurate prediction for heart, kidney, and diabetes diseases. With the stochastic gradient decent algorithm we learn our model effectively and get the better result for the heart, kidney, and diabetes disease. After the learning we can produce the confusion matrix for our model that gives the prediction matrix. Based on the confusion matrix we check the accuracy of our model.

Step 4- Deployment and analysis on real-life scenario of peoples

By appalling the prediction model does the diseases prediction effectively. Trained datasets and tested data sets are deployed in real-life scenario made by human experts and will leverage for future improvement.Figure2 (a) and (b) show the overall flow of the project. Methodology follows the following architecture as,

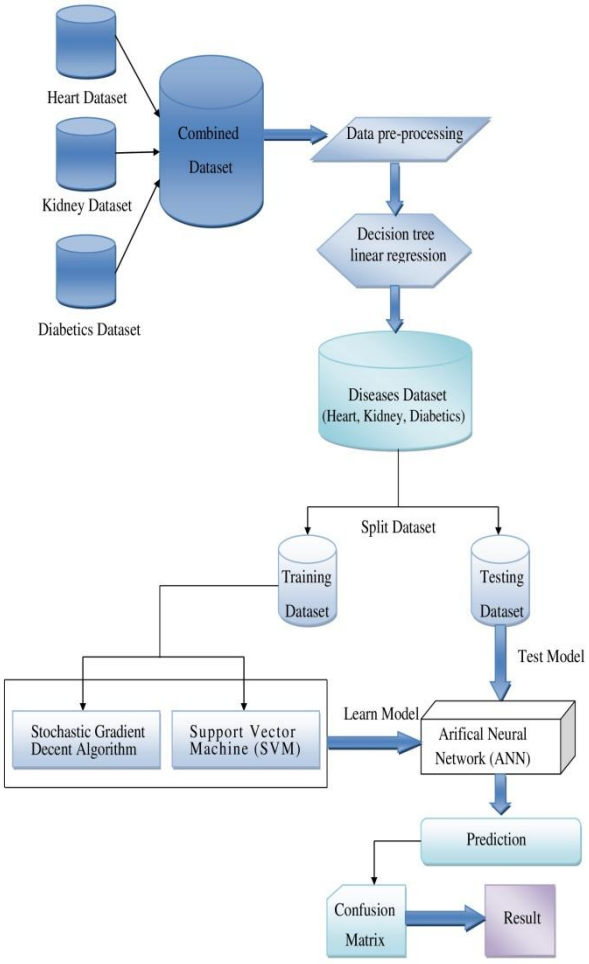

Figure: 2 (a)

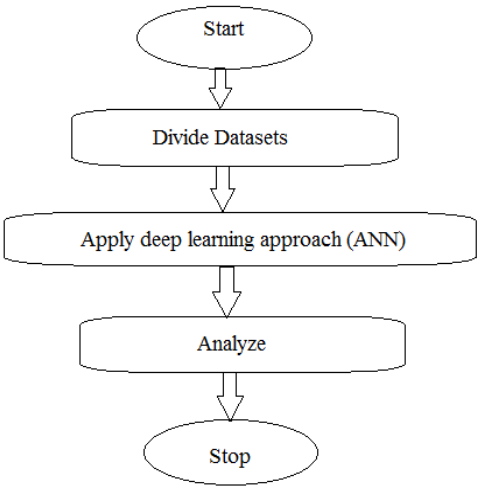

Figure: 2 (b)

Figure 2 (a) & (b) Workflow of diseases prediction system using deep learning based on treatment history and health data.

4. EXPERIMENTAL RESULT

For performance evaluation of the experimental result we consider the TP, TN, FP, and FN as the true positive (correctly predictions), true negative (correctly rejected predictions), false positive (incorrect predictions), and false negative (incorrectly rejected predictions), respectively. For the experimental result we obtain the accuracy, precision, recall, and f1-score as follow,

$$\textbf{Accuracy} = \frac{\text{TN/TP}}{\text{TN+FP+FN+TP}}$$

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\textbf{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\textbf{F1 - Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision+recall}}$$

Accuracy is ratio of the (TP/TP) / (TP+FP+FN+TP) means total no of correct prediction divided total number of dataset. Precision is the overall the ratio of the TP/ (TP+FP). In the same way we calculate the recall as the ratio of TP/ (TP+FN). F1-score is the harmonic mean which is obtained from the precision and the recall.

In this section we discuss about the experimental result of the model. For doing the experimentation we contain the 1471 no of records of dataset. In the 1471 records we split the dataset into the training and testing datasets. For the training dataset use the 1176 records and for the testing purpose use the 295 records out of 1471. For experimental result we do the comparison between the two models as SVM and the ANN. In that the experimentation is done on training and testing datasets. When doing the prediction with the SVM model got the less accuracy as compare to ANN model. With SVM model and ANN model we predict the diseases effectively. Then obtain the precision, recall, and f1-score for the actual class input as 0 and 1. In that the 0 is used as no heart diseases, no kidney diseases, and no diabetes diseases. 1is used as the heart diseases, kidney disease, and diabetes diseases.

With the ANN model for heart diseases, kidney diseases, and diabetes diseases got the 95%, 98%, 72% accuracy respectively. The figure 3 shows the precision, recall, and f1-score for heart, kidney, and diabetes diseases. In that for heart diseases got 95%, 95%, 95% precision, recall, and f1-score with the 137 and 158 for no heart diseases and heart diseases out of the 295 occurrences of records.

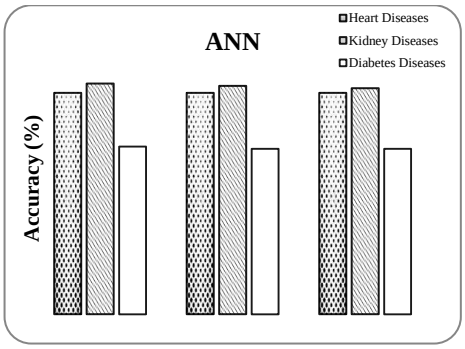

Figure 3.ANN model precision, recall, f1- score for heart, kidney, and diabetes diseases.

For the kidney diseases got 99%, 98%, 97% precision, recall, and f1-score with 42 and 253 for no kidney diseases and kidney diseases no of occurrences respectively. Also for the diabetes diseases got 21%, 71%, 71% got precision, recall, and f1-score for no diabetes diseases and diabetes diseases with 163 1nd 132 no of occurrences.

With the SVM model prediction got the 79%, 93%, 71% accuracy for the heart disease, kidney disease, diabetes diseases respectively. Figure 4 shows the precision, recall, and f1-score for heart, kidney, and diabetes diseases.

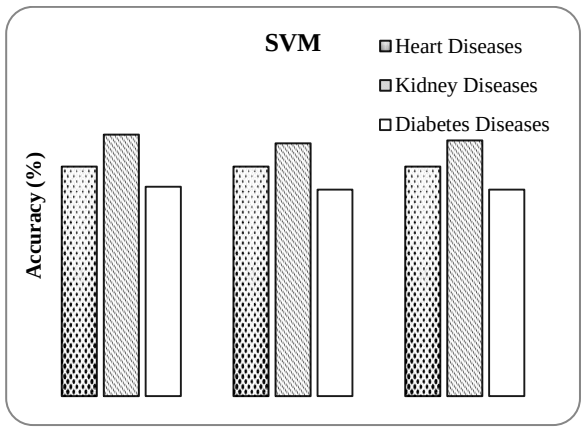

Figure 4.SVM model precision, recall, f1- score for heart, kidney, and diabetes diseases

In that for heart diseases got the 79%, 79%, 79% precision, recall, and f1-score for 139 and 156 no of occurrences for no heart diseases and heart diseases. Similarly for the kidney diseases 90%, 87%, 88% of precision, recall, and f1-score are got for 47 no occurrences of no kidney disease and 248 no of occurrences as kidney diseases. Also for diabetes diseases got the precision, recall, and f1-score as 72%, 71%, and 71% for 162 as no diabetes diseases and 133 numbers of occurrences for diabetes diseases.

As shown in the figure 5(a), for ANN and SVM model average precision accuracy in the percentage. For ANN model got the average precision accuracy as 95%, 98.5%, 72% for heart, kidney, and diabetes diseases and for the SVM model got 79%, 87%, 71% average precision accuracy for heart, kidney, and diabetes diseases.

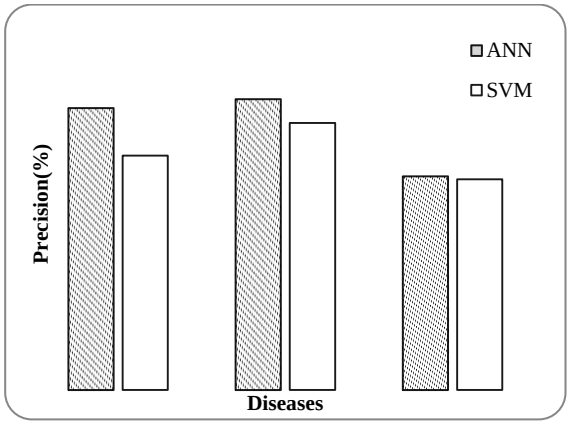

Figure 5(a) Average precision accuracy of ANN and SVM for heart, kidney, and diabetes diseases.

Figure 5(b) shows the average recall accuracy in the percentage for ANN and SVM model. With ANN model got the 95%, 98%, 70.5% accuracy for heart, kidney, and diabetes diseases. With SVM model got average recall for heart, kidney, and diabetes as 79%, 87%, and 71% respectively.

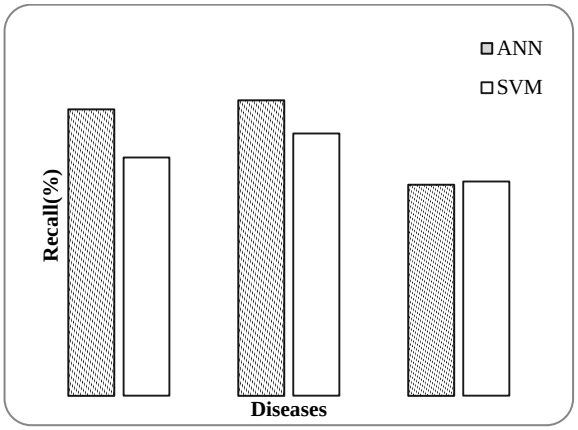

Figure 5(b) Average recall accuracy of ANN and SVM for heart, kidney, and diabetes diseases

5. DISCUSSION AND ANALYSIS

In that we discuss the overall performance of the model. For the experimentation we use the 1471 no of records of dataset. In that we split the dataset into the training and testing datasets. For the training gives the 1176 records and for the testing we give the 295 records out of 1471. We build the model by using ANN. In that we compile the model and fit the model with the 100 no iteration, and 70 batch size. In that we calculate the accuracy and find the loss of model.

In this section we compare the result of ANN model with the traditional SVM method. With the ANN model we got the more accuracy as compare to the SVM model. Figure 6 shows the overall accuracy of ANN and SVM model for heart, kidney, and diabetes disease. With the SVM model got the accuracy as 79%, 93% and 71% for the heart, kidney, and diabetes diseases respectively. But with the ANN model got the higher accuracy than SVM model as 95%, 98% and 72% for heart, kidney and diabetes diseases.

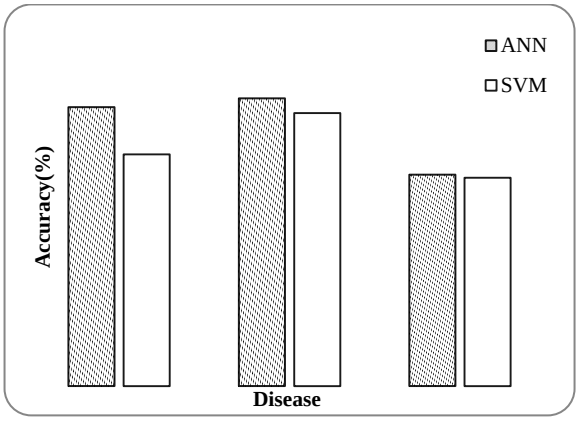

Figure 6.Overall accuracy of ANN and SVM for heart, kidney, and diabetes diseases.

Figure7 shows accuracy and loss for the heart, kidney, and diabetes diseases. For the heart, kidney, and diabetes disease train ANN model with 100 number of iteration. In that when training is done with model for above diseases we observed that training loss is decreases and validation loss is increases for all three diseases.

Figure 7(a) shows heart disease training loss vs. validation loss and training accuracy vs. validation accuracy. For heart disease 0.006% training loss and 0.25% validation loss, 0.99% training accuracy and 0.95% validation accuracy are got for 100 number of iteration.

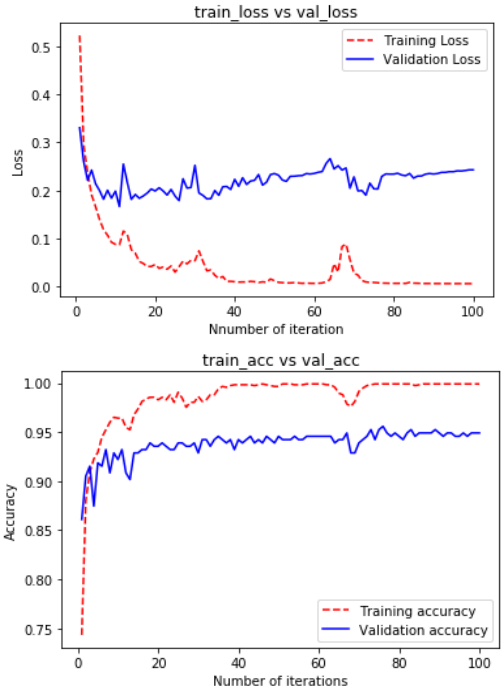

Figure7 (a) Training loss vs. validation loss AND training accuracy vs. validation accuracy of heart diseases

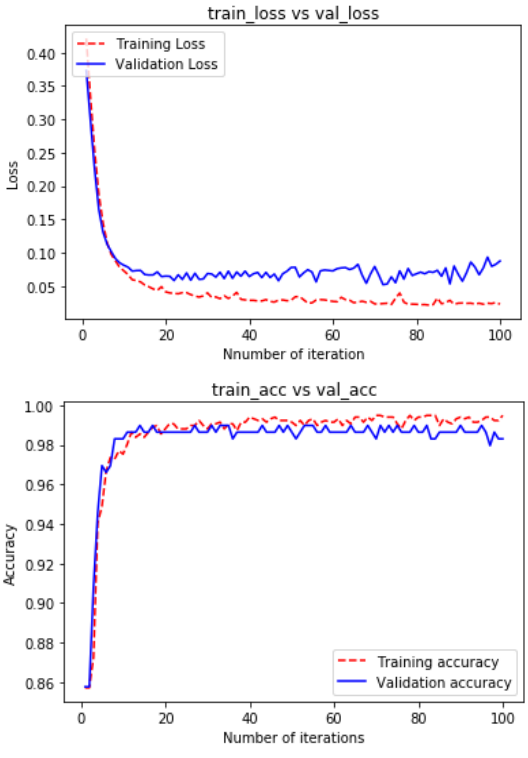

Figure7 (b) Training loss vs. validation loss AND training accuracy vs. validation accuracy of kidney diseases

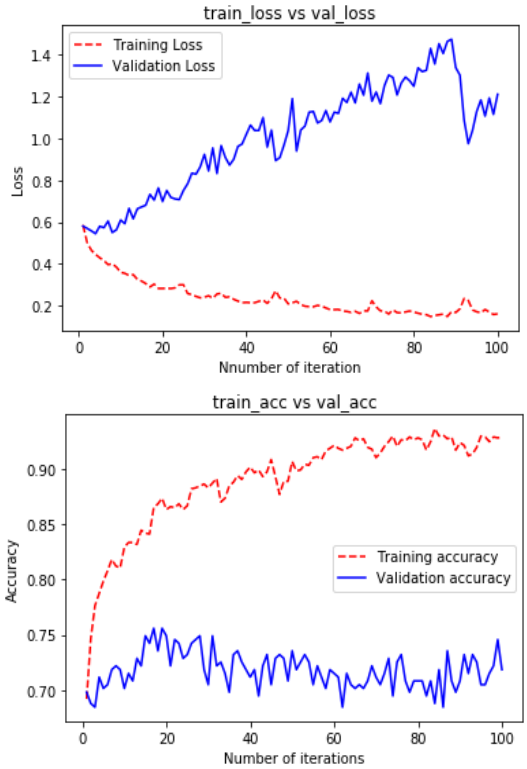

Figure7(c) Training loss vs. validation loss AND training accuracy vs. validation accuracy of diabetes diseases

Figure7. Training loss vs. validation loss and Training accuracy vs. validation accuracy for heart, kidney and diabetes diseases.

Figure 7(b) shows kidney diseases training loss vs. validation loss and training accuracy vs. validation accuracy. For kidney disease got 0.038% as training loss and 0.067% as validation loss and 0.98% training accuracy and 0.98% validation accuracy for 100 number of iteration.

Figure 7(c) shows diabetes disease training loss vs. validation loss and training accuracy vs. validation accuracy. For diabetes diseases got the 0.13%, 0.066% loss and 94%, 72% accuracy for 100 numbers of iterations.

5. CONCLUSION

The aim of the study was on disease prediction in humans based on historical data. In this paper, a new Probabilistic model and deep learning method is proposed. In that pre-processing is applied on dataset and remove all the unnecessary data and unwanted data. Then extract important features from that data, it runs effectively on healthcare databases. By using the ANN we build the model. Model is trained by using the scholastic gradient decent algorithm, and tested model by using the artificial neural network. By using that model we got the prediction accurately and produce Confusion matrix for that data. Our model gives the 95%, 98%, 72% accuracy for predicting the heart, kidney, and diabetes diseases. That model will be deployed in a real-life for the predicting the diseases. By early predicting the

diseases using the proposed system reduce the cost of treatments because previous data of patient is available to doctors and they do the prediction of diseases based on that data. Disease can be predicted than early treatment which can reduce the risk of patient life. Because of the deep learning algorithm they give high accuracy and take less time.

ACKNOWLEDGEMENT

We are thankful to the center of excellence AL and DL CMRCET for the support in the execution of the work. We especially thankful to Mr. S. Siva Skanda Associate Professor and Miss. A. Poongodai Associate Professor of CSE for giving the guidance when-ever needed.

REFERENCES

1. Min Chen, Kai Hwang, Yixue Hao, Fellow, Lu Wang, and Lin Wang," By using Machine learning Algorithm predicting the diseases with Big Data technology from the Healthcare Commu-nities," IEEE, 2016.
2. R.Tamilarasi, Dr.R.Porkodi Department of Computer Science Bharathiar University, Coimbatore, Tamil Nadu, India "A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare", International Journal of Emerging Research in Management and Technology ISSN: 2278-9359 (Volume-4, Issue-3) (IJERMT) , March 2015.
3. Darcy A., Davis, Nicholas Blumm, Nitesh V.Chawla, "Pre-dicting Individual Disease Risk Based on Medical History", ACM, 2008.
4. Feixiang Huang, Chien-Chung Chan, and Shengyong Wang "Predicting Disease By Using Data Mining Based on Healthcare Information System", IEEE International Conference on Granular Computing, 2012.
5. Abhishek Rairikar, Vikas Sabale, Vedant Kulkarni, Harshavardhan Kale "Heart Disease Prediction using Data Mining Techniques", International Conference on Intelligent Computing and Control (I2C2), 2017.
6. Saurabh Pandey, Manish Madhava Tripathi "Diagnosis of Diabetes using Artificial Intelligence Techniques by using Bio Medical Signal Data", International Journals of Research and Development in Applied Science and Engineering ISSN: 2454-6844 (IJRDASE), 2017.
7. Dr.B.Srinivasan, K.Pavya,"A Study for Prediction in Healthcare Sector using Data Mining Techniques", International Research Journal of Engineering and Technology Volume: 03 Issue: 03 (IRJET) March, 2016.
8. Paritosh Khubchandani, Suraj Lala, Kundan Jha, Pallavi Saindane, Rohit Bijani, "Medical diseases Prediction using Arti-ficial Intelligence techniques", International Journal of Engineering Science and Computing (IJESC) March, 2017.
9. Smitha.T, V.Sundaram "An Efficient Algorithm for Disease Prediction with Multi Dimensional Data", International Journal of Computer Application (0975-8887) Volume 63-No.9 (IJCA), February 2013.
10. Anandanadarajah Nishanth, Tharmarajah Thiruvaran," Iden-tify the important attributes of chronic kidney diseases for early detection", IEEE, 2017.
11. Fatma Taher, Hussain Al-Ahmad, Naoufel Werghi, Rachid Sammouda "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods", IEEE 2011.
12. Ravi Sanakal, S.T Jayakumar, "Use of Data mining algo-rithm Fuzzy C Means Clustering and Support Vector Machine for Prognosis of Diabetes", International Journal of Computer Trends and Technology (IJCTT), Volume – 11, 2014.
13. M. H. Muhamad Adnan, W. Husain, N. A. Abdul Rashid, "Data Mining techniques for Medical Systems", International Conferences on Advances in Computer and Information Tech-nology (ICACIT), 2012.
14. L.G.Kabari, E.O.Nwachukwu, "Eye diseases diagnosis by sing the neural network and decision tree algorithm", INTECH, 2012.
15. Loredana Stanciu, Adriana Albu, "Artificial Intelligence Benefits in Medical Predictions", November, 2015.
16. P. H. Winston, J. Beal, "Introduction of the New Frontier of Human Level Artificial Intelligence", Intelligence system, IEEE, 2009.
17. Qeethara Kadhim, Al-Shayea, Itedal S. H. Bahia(2010), "By using Artificial Neural Networks approach diagnosis the Urinary System Diseases", IJCSNS International Journal of Computer Science and Network security, Vol.10 No.7.
18. Margaret H. Dunham, "Data Mining Introductory and Ad-vanced Topics of data mining techniques", Prentice Hall, 2003.
19. M.Haitham, A.V.Sahakian, A.Angari, "By using the Support Vector Machine (SVM) classifier recognition a Sleep Apnea Syn-drome", IEEE Volume – 16, May 2012.
20. G.E.Sakr, H.A.Huijer, I.H.Elhajj, "Support vector machines used for defining and detecting Agitation Transition", IEEE, Vo-lume - 1(pp. 98-108), December 2010.
21. Haya Alasker, Shatha Alharkan, Lala Septem Riza Wejdan Alharkan, "Intelligent Classifiers used for detect the kidney dis-ease", International conference on Science in Information Tech-nology (ICSIT) 2017.
22. Qeethara Kadhim , Al-Shayea and Itedal S. H. Bahia(2010), "By using Artificial Neural Networks approach diagnosis the Urinary System Diseases", IJCSNS International Journal of Computer Science and Network security, Vol.10 No.7.
23. Mohammad Taha Khan, Dr.Shamimul Qamar, Laurent F. Massin, "Using the data mining techniques predicts the cancer and heart diseases", International Journal of Applied Engineering Re-search (IJAER), 2012.
24. F.Azuaje, K.Adamsom, W.Dubitzky, P. Lopes, N. Black, "Predicting Coronary Disease Risk by using Neural Network Approach Based on Short-term Interval Measurements ", Artifi-cial Intelligence in Medicine, Volume – 15,March 1999.
25. Anunciacao Orlando, Gaspar Jorge, Oliveira L.Arlindo Gomes C. Bruno, Vinga Susana, "A Data Mining approach for detecting the Breast Cancer disease," Advances in Soft Computing (ASC), volume-74, 2010.
26. Ma.jabbar, B.L.Deekshatulu, Dr.prirti Chandra, "Heart attack prediction using Association rule mining with clusters", Journal of Theoretical and Applied Information Technology (JTAIT), 2011.
27. Veenita Kunwar, A. Sai Sabitha, Khushboo Chandel, Abhay Bansal, "The data mining techniquesused for analysis chronic kidney diseases", International conference, 2016.
28. Sellappan Palaniappan, Rafiah Awang,"With the data mining techniques predicting the intelligent heart diseases", International Conference on Computer Systems and Applications (ICCSA), April 2008.
29. Dr.Y.S.Kumaraswamy, Shantakumar B.Patil, "Heart attack diseases prediction extract the patterns from the heart diseases warehouses", International Journal of Computer Science and Net-work 228 Security (IJCSNS), 2009