ii LITERATURE REVIEW

In this section different research works that were envisioned to predict the risk of stroke and other diseases.

*A.* STROKE RISK PREDICTION THROUGH NONLINEAR SUPPORT VECTOR CLASSIFICATIONS MODELS

In this research paper Sabibullah Mohammaed Hanifa and Kasmir Raja S.V presented a study to find the possible risk of stroke by subjecting the risk factors to Support Vector Machines. The authors used the SVMLight software for implementation. Their data set was extracted from cohort of population set of various hospital situated at Tiruchirappali city, Tamilnadu, India. They used 100 patient’s data with 8 attribute called Hypertension, Diabetes Mellitus, Obesity, Cigarette Smoking, Heart Disease, Prior Stroke, High Cholesterol, and Physical Activity. They used Support Vector Classification model parameters through its kernel function named polynomial kernel and Gaussian (RBF) kernel. The authors evaluated the result through Confusion matrix and show that the rate of correctness of prediction by RBF is 98% where by polynomial is 92%. So the author told in this paper that the application of SVM models can be used for the processing of stroke related risk factor data.

*B.* DIABETES PREDICTION USING MEDICAL DATA

In this research paper the author presented a diabetes prediction system to diagnosis diabetes. They tried to improve the accuracy in diabetes prediction using medical data with various supervised machine learning algorithms namely Naïve Bayes (NB), Multilayer-perceptron (MLP), Random Forest (RF) and the accuracy is noted with different test methods such as 10-fold cross validation (FCV), use percentage split with 66%(PS), and using training dataset (UTD) with pre-processing method and without pre-processing methods. For this work the author used WEKA software tool. They used PID dataset and collected dataset with the medical report of 768 persons include 8 features. However, the concluded that the pre-processing methods increase accuracy for Naive Bayes algorithm.

*C.* A REVIEW ON PREDICTION OF MULTIPLE DISEASES AND PERFORMANCE ANALYSIS USING DATA MINING AND VISUALAIZATION TECHINIQUES

The paper was motivated to construct a basic prototype model which can determine unknown knowledge related with multiple disease from past database records of specified multiple diseases. In this research work, at first the authors experimented on the dataset consisting of 1000 records and 14 attribute and the Naive Bayes algorithm gave better accuracy than others. Then the authors experimented again on the dataset of 1000 records within 76 attributes and back-propagation algorithm gave the best accuracy (100%) where the accuracy is given by Naive Bayes algorithms is 90.74%. However, the authors told in their paper that the accuracy result can be changed if we increase the number of datasets and the number of attributes.

*D.* BREAST CANCER CLASSIFICATIONS USING SUPPORT VECTOR MACHINE AND NEURAL NETWORK

In this work, the author tried to carry out using Wisconsin Diagnosis breast Cancer database to classify the breast cancer as either benign or malignant. They used dataset consists of 400 observations of patients. Among them 300 are benign and 100 are malignant status. Each instance has 20 features. They used two classes for classifications named cancerous cell and non-cancerous cell and the experiment was carried out by using SVM and Neural Network. The result both SVM and Neural Network were compared on the basis of accuracy and precision. For this experiment, NN technique is more efficient compare to SVM technique. However, the author observed that NN technique is more efficient but the difference is very low.

*//E.* DIAGNOSIS OF DIABETES USING CLASSIFICATION

MINING TECHNIQUES

The authors have employed two algorithms namely J48 and Naïve Bayes to create the model for diagnosis in determining diabetes in women. They used PID (Pima Indians Diabetes) database of National Institute of diabetes and Digestive and Kidney Diseases dataset pre-processed in CSV format as their input. The data was divided into training set and test set by 10-fold cross validation and percentage split techniques. They pre-processed their dataset by using Weka software tools. The J48 algorithm is used on the dataset using WEKA after which data are divided into “tested-positive” or “tested-negative” depending on the final result of the decision tree that is constructed.

*F.* APPLICATIONS OF MACHINE LEARNING IN CANCER PREDICTION AND PROGNOSIS

In this paper, the author have described how the machine learning works well in cancer prediction and prognosis. They have been showed two histograms. The first one is about the steady increase in published papers using machine learning to predict cancer risk, recurrence and outcome. The another one is about the frequency with which different types of machine methods namely: Naïve Bayes, Genetic algorithm, Fuzzy logic, Clustering, SVM, ANN, Decision Tree are used to predict different types of cancer such as bladder, breast, colorectal, liver, lung, lymphoma, Prostate, Skin, Throat etc. They discussed in their paper about 3 case study which was about prediction of cancer risk or susceptibility, survivability, recurrence. They worked very well. However, all machine learning studies are not conducted with the same attention by them.

*//*LIVER DISEASE PREDICTION USING SVM AND NAÏVE BAYES ALGORITHMS

The author have been used classification algorithms namely Naïve Bayes and SVM for liver diseases prediction in this research work and they implemented their work in Matlab 2013 tool. The authors described the classification of liver diseases such as cirrhosis, bile duct, chronic hepatitis, liver cancer and acute hepatitis. In this work paper, they showed that SVM algorithms works better than Naïve Bayes algorithms on the basis of performance accuracy. But if we think about execution time, then Naïve Bayes classifier needs minimum execution time.

*//H*. A BRIEF SURVEY ON THE TECHNIQUES USED FOR TECHNIQUIES THE DIAGNOSIS OF DIABETES MELLITUS

This paper described about different data mining methods such as k-ford cross validation and classification, Classwise K Nearest Neighbor ,SVM, LDA-Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, Statistical Normalization, Back Propagation. At first the authors have been collected 768 cases but after deleting the missing values they had 460 cases for their experiment. They had compared the accuracy of the performance of those classification algorithms and observed that the SVM gave best accuracy as 81.77% compare to others.

// *H.* PREDICTION OF STROKE THROUGH STACKED TOPOLOGY OF ANN MODELS

This paper was motivated to predict the stroke risk by proposing the stacked ANN topology model with higher prediction accuracy. They have collected 300 data for their paper from different hospitals at Tiruchirappalli city, Tamilnadu, India. They analysed their data by using back propagation algorithm and implemented those data through MATLAB 7.3.0-Neural Network Toolbox. They divided their output in three categories namely: High risk, Moderate risk and Low risk. They used ANN model consist of three layers, respectively input layer, output layer and one internal layer. They used Confusion matrix as a method for finding error. They also presented a graphical structure of their predicted and actual outputs of all network. They have been presented a good paper. However, sometimes they made a big difference between their predicted outputs and actual outputs.

//PREDICTION AND CONTROL OF STROKE BY DATA MINING

This paper described about stroke and the risk factors of stroke. The authors collected 807 data sets within 50 risk factors for stroke by using a standard checklist during year 2010-2011 in Iran. After pre-processing and cleaning data they have used WEKA Software tool for implementation and data mining techniques such as K-nearest neighbor and C4.5 decision tree for analysing the data sets. In this work they have found the performance accuracy of the C4.5 decision tree was 95.42% and the K-nearest neighbor was 94.18%. So they told C4.5 decision tree and K-nearest neighbor can be use for prediction of stroke. However, the authors presented a description about C4.5 decision tree and K-nearest neighbors only.

// ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION

Antonis Lambrou, Harris Papadopoulos et.al have been presented a research paper to provide reliable confidence measures for the assessment of stroke by using the Conformal Prediction framework. They evaluated their results of four different conformal prediction respectively: ANN, NB, SVM and K-NN. For experimentation the authors applied Principal Component Analysis (PCA) on the dataset and have selected its 6 features which accounted of 98% of its variance. They have used Leave-One-Out (LOO) method for evaluating. The classifier algorithm ANN-CP was structured with one hidden layer consist of 3 units and the output layer consist of 2 units. In this paper the authors observed that SVM classifier has the best accuracy. However, when the authors increase the percentage of confidence the certainty rates start to decrease.

// AN INTEGRATED MACHINE LEARNING APPROACH TO STROKE PREDICTION

Aditya Khosla, Yu Cao et.al. published a paper to presented an integrated machine learning approach combining the elements of data imputation, feature selection and prediction. They used different metrics for evaluating their methods such as Notation, Area under the ROC curve (AUC), Concordance Index. They handled their missing data by using mean, median, imputation through linear regression. They evaluated their data imputation quality by using 10-fold cross validation and used SVM for stroke prediction.

The showed in their paper that the combination of Conservative Mean feature selection and Margin-based censored regression gave the best performance.

//PRIEDICTION OF STROKE USING DATA MINING CLASSIFICATION OF STROKE

This research was carried out by Ohoud Almadami and Riyad Alshammari to predict patient at risk of developing stroke by using data mining techniques and to find the patient with who has higher chances to develop stroke. They used three classifier algorithms namely: C4.5, Jrip and multi layers perceptron (MLP). They collected 969 data sets from National Guard hospitals in three different cities in Kingdom of Saudi Arabia. They collected their data in 2016 from 2nd January to 31st September. They divided their data sets into two classes. First one includes with the patient who have stroke and the 2nd class includes the patient who has mimic stroke but they have diagnosis as they have stroke. The authors have made a train data set to build their model with 10-fold cross validation and have made a test data set to evaluate the model. They have used WEKA Software tool for applying their data mining techniques. In this research work it is observed that with the comparison of 10-fold cross validation the Jrip classifier algorithm gave the best performance accuracy (92.60%) but after applying PCA on stroke data C4.5 algorithm gave the best performance on test data set (95.25%).

//A MODEL FOR PREDICTING ISCHEMIC STROKE USING DATA MINING ALGORITHMS.

Balar Khalid and Naji Abdelwahab have done a thesis paper to provide a model for predicting Ischemic stroke using data mining algorithms. They have emphasised on prediction and finding risk factors for ischemic stroke so that they could provide a model. They analysed their data sets through C4.5 DT algorithm and Logistic regression and WEKA 3.6 Software tool. They analysed their data by using a software of Microsoft “XLSTAT”. This software based on Visual Basic language. They have found the rates of sensitivity was 77.58% and specificity was 83.03% and error rate was 19.7% with this “XLSTAT” software and AUC (Area Under Curve) area under ROC (Receiver Operating Curve) equals 0.89.