A Brief Overview of Data Mining Technologies for Diabetes and Diabetes Risk Prediction

Abstract—**Diabetes is one of the most life threatening diseases that may affect more than millions of people by 2030 including the youngster. Unlike flu or cold diabetes doesn't have specific symptoms. But the general symptoms or medical data are helpful to predict diabetes risk or diabetes at its early stage, which can reduce a sudden death because of uncontrolled diabetes. The pre-eminent role of data mining approaches for disease prediction is trending towards appropriate and cost-effective treatment. Multiple data mining techniques namely Naïve Bayes (NB), Decision tree (DT),Logical regression(LR) Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like Kth-Nearest Neighbor (KNN), and Support vector machine (SVM) algorithms have been well-accepted by the researchers for diabetes prediction over the recent years. In this paper, we are aimed to provide a comparatively quick and easy review of available data mining technologies that have been applied for diabetes prediction in existing research works. The comparative study of the related works exhibits the data mining approaches and obtained model accuracy of different works of literature. A detailed study on the data mining classification techniques, evaluation and testing models and mining tools have been presented for introducing new research activities appropriately for diabetes risk prediction at the early stage.**

Index Terms—**Diabetes Risk, Data Mining, KDD, Dataset, Evaluation Model**

I. Introduction

Diabetes is one of the most dangerous and deadliest diseases in the current world. According to the world health organization(WHO), the number of people with diabetes was 108 million in 1980 which has risen to 422 million in 2014 [1]. According to WHO's Fact Sheet, 2012, about three out of four people with diabetes are from the low and middle-income country. Diabetes is known as the mother of other infectious diseases like tuberculosis(TB), malaria and HIV/AIDS. About 15 percent of TB is thought to be due to diabetes. In South-East Asia Region nearly 1 million people die from consequences of high blood sugar every year [6]. It is predicted that there were 175 million diabetics in 2000 and the number will increase to 354 million by 2030. Also said that 85 percent of the world's diabetes patients will be in developing countries.

Diabetes is a serious, lifelong situation where your blood glucose level above normal [2]. Generally, we know about two types of diabetes called type 1 and type 2 diabetes. Type 1 diabetes occurs when the immune system mistakenly attacks and very little insulin released to the body or sometimes no insulin released to the body. Type 2 diabetes occurs when our body doesn't produce proper insulin or the body becomes insulin resistant. Some researchers divided diabetes into Type 1, Type 2, and gestational diabetes [6]. Gestational diabetes is a type of diabetes which occurs only in pregnancy due to hormonal change. There is a fourth kind of diabetes that was added by other researchers known as pregestational diabetes. Pre-gestational diabetes occurs when one has type 1 or type 2 diabetes before becoming pregnant. Recently a diabetes specialist divided it into six types for proper diagnosis called (i)type 1,(ii) type 2, (iii) gestational, (iv) LADA(Latent Autoimmune Diabetes of Adulthood), (v) MODY(Maturity Onset Diabetes of the Young), (vi) NDM(Neonatal Diabetes Mellitus) [3]. The common symptoms of diabetes are polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity etc.

Awareness about the disease and treatment for those who are affected or have a risk is very poor. About half of the affected people don't even know that they have diabetes and a very small percentage of them get treated. Currently, there is no cure for this deadly disease. But, if one can detect it in an early stage, he can lead a healthy life. Early detection also helps to reduce the risk of serious complications such as stroke, premature heart disease, blindness, limb amputations, and kidney failure [7].

A doctor can identify a disease by listening to symptoms and different physical tests, for an example, if a patient comes with the symptoms like polyuria, polydipsia, sudden weight loss or some others, then the doctor can predict from his knowledge that, it might be diabetes. However, for being surer he or she measures the blood sugar level to identify diabetes. It is a long process, which costs patients more time and money. In this modern era of technology, computer technology can help us to detect diseases accurately and can save our time and money. Data mining is an important field of computer science which used for prediction. It is the process of discovering new data from previously known data through data analysis [3].To predict a disease using data mining approaches we need its symptoms along with clinical data. Symptoms are a very important factor for new patients and early stage prediction since they have no data except symptoms. We also need clinical data for analyzing and discovering new data.

Discovering new data is a process which converts low-level data into high-level knowledge, in term of computer technology we called it Knowledge Discovery from Data(KDD). KDD works with 4 steps: selection, preprocessing, transformation, data mining and evaluation [4]. There are different classification approaches to predict disease with data mining such as classification, clustering, and association. Decision tree induction, Naive Bayes classifier, k-nearest neighbor, genetic algorithms, support vector machine, fuzzy set are the algorithms for classification techniques. Association rule is used for descriptive patterns. K-means, K-medoids, chameleon, statistical and neural network approaches are the clustering algorithms [8]. Many researchers have done their researches with different data mining processes obtained different accuracy. Accuracy found by different data mining algorithm may differ because of the different dataset, parameters, and condition of data. One algorithm's accuracy may better for a particular algorithm while another can have a bad accuracy result.

Since different researchers research with different datasets, attributes, and different DM algorithms, we need an organized direction to predict diabetes risk in its early stage. So in this paper, we present an empirical study of the popular data mining approaches, evaluation model, and dataset testing tools. It also represents the different methodology and, achieved accuracy by different researchers in a convenient way. Finally, this paper proposed its future scope to predict early-stage diabetes risk using patients symptoms with the help of data mining techniques.

The rest of the paper is organized as follows: Section II provides the literature review with a comparative table of analysis. Section III describes the detailed study of the background of data mining for diabetes prediction. In section IV we identified the problems and future work direction of this research Finally, in Section V, we conclude the paper.

II. Literature Review

In this section different research works that were envisioned to predict diabetes using data mining approaches are provided. A comparative information analysis of the mostly related works are represented in Table I.

A. A DATA MINING APPROACH FOR PREDICTION AND TREATMENT OF DIABETES DISEASE [8]

This paper uses different data mining algorithms like Naive Bayes, J48(C4.5) JRip , Neural networks, Decision trees, KNN, Fuzzy logic and Genetic Algorithm. They collected 865 data with 9 attributes called Sex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 hour Serum Insulin and Diabetes probability and used WEKA 3.6.6 for the experiment. They find 100% accuracy with J48(C4.5), 98.48% with the Decision Tree, 97.85% with the Neural Network, 96.54% with JRip and 95.85% with Naive Bayes algorithm. They also calculate the performance over time. Here they find 68.58% accuracy with J48(C4.5) and it took 658 minutes, 52.58% with the Decision Tree and it took 875 minutes, 50.68% with the KNN and it took 956 minutes, 65.48% with JRip and it took 765 minutes and 55.85% with Naive Bayes algorithm which took 845 minutes.

B. A COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR PREDICTION OF POSTPRANDIAL BLOOD GLUCOSE : A COHORT STUDY [9]

In this paper, the authors used 5 data mining algorithms including RF, C4.5, SVM, MLP, and LR to predict postprandial blood glucose. They wanted to compare the results of these algorithms. Their dataset was collected from LandSeed Hospital from the time 2006-2013. They checked the performance of the algorithms using WEKA tool using 10-fold cross-validation method. They experimented the correct classification rate (CCR), sensitivity, and sensitivity of all the five algorithms and created a comparative table. Their CCR results of RF,C4.5,MLP,LR,SVM are 82.68%,76.56%,75.53%,74.48%,69.61% sequentially. Finally, they found the RF model as best for the prediction for postprandial blood glucose.

C. A BRIEF SURVEY ON THE TECHNIQUES USED FOR THE DIAGNOSIS OF DIABETES-MELLITUS [10]

In this work, the author discussed different data mining techniques and software tools which are used for predict diabetes. They used 738 patient's data for experimental analysis. After the deletion of missing values, they had 438 data remaining. To predict diabetes they introduced heterogeneous data mining algorithms like CNN, KNN, SVM, SVM+LDA, NB, SVM, ID3, C4.5, CART for comparing the analysis on the dataset. The best accuracy at 88.10% was achieved using SVM and LDA algorithm together.

D. DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES [11]

The research was carried out by Tejas N. Joshi, Prof. Pramila M. Chawan. They compared three machine learning algorithms to predict diabetes. They introduced SVM, Logistic regression, ANN to seven attributes of their data including the Glucose, Blood Pressure, Skin, Thickness, Insulin, BMI, Diabetes Pedigree Function, and the age. After comparing their features, the researcher opinionated that the Support Vector Machine(SVM) is best for diabetes prediction considering the specified attributes of a certain dataset. However, this prediction mechanism requires a large amount of data.

E. PREDICTION OF DIABETES BY USING ARTIFICIAL NEURAL NETWORK [12]

This paper used Artificial Neural Network for predicting diabetes. They collected 250 diabetes patients data from Pusat Perubatan University Kebangsaan Malaysia, Kuala Lumpur and between 25 to 78 years old. They used MATLAB to train data. They had done Regression analysis using different algorithms, BFGS Quasi-Newton, Bayesian Regulation, Levenberg-Marquardt. They found 88.8% accuracy with Bayesian Regulation algorithm.

F. DIABETES PREDICTION USING MEDICAL DATA [13]

This paper was motivated to improve the accuracy of diabetes prediction using different machine learning algorithms. They also used PID dataset for prediction. They used WEKA as their software tool. They tested their dataset with Naive Bayes(NB), Random Forest(RF), and function-based multilayer perceptron (MLP) algorithms and used different test methods called FCV, PS, UTD. They also predict with pre-processed and without pre-processed data and made a convenient table on their result. They found 100% accuracy with Random Forest algorithm with UTD method. However, the author concluded that pre-processed data gives more accuracy in the Naive Bayes algorithm

G. USING DATA MINING TO DEVELOP MODEL FOR CLASSIFYING DIABETIC PATIENT CONTROL

LEVEL BASED ON HISTORICAL MEDICAL RECORDS [14] Tarig Mohamed Ahmed has done a research for developing a model using data mining for diabetic patients. This paper divided their patient's data into two class: under control (HbA1c < 7%) and out of control (HbA1c >7%). They used three algorithms of data mining called Naive Bayes, Logistic Regression, and J48. They used WEKA to implement the model. They find the best accuracy 74.8% using the Logistic algorithm.

H. DEVELOPING A PREDICTED MODEL FOR DIABETES TYPE 2 TREATMENT PLANS BY USING DATA MINING [15]

In this paper, the researcher Tarig Mohamed Ahmed has created a new model for type 2 diabetes patients treatment. He collected 318 medical records with 9 nominal attributes including the patient's Gender, Age, Smoking, History of hypertension, Renal problem, Cardiac problem, Eye problem. Duration of Diabetes Basic control was used as a class level attribute. He used the J48 algorithm and found an accuracy rate of 70.8% and ROC (Receiver operating characteristic) rate was 0.624.

I. DIABETES PREDICTION BY SUPERVISED AND UNSUPERVISED LEARNING WITH FEATURE SELECTION [16]

Rabina, and Er. Anshu Chopra has done a thesis paper to predict diabetes with supervised and unsupervised learning. They used the software tool WEKA to find a better prediction algorithm in machine learning. Finally, they concluded that ANN and Decision tree are best for diabetes prediction.

J. ANALYZE DATA MINING ALGORITHMS FOR PREDICTION OF DIABETES [17]

In this paper, the author presented a study about some data mining algorithms for disease prediction like Gaussian Naive Bayes, KNN, SVM, and Decision Tree. These algorithms can be used to predict diabetes. They used Pima Indian Diabetic Set from the University of California, Irvine (UCI) Repository of Machine Learning Databases. They discussed the background study of data mining in an easy and understandable way. They found 69.685% accuracy with .33 error rate using the Naive Bayes algorithm, 70.866% accuracy, and .34 error rate while introducing KNN, and 64.174% accuracy with .29 error rate for the Decision tree. However, with the KNN algorithm, the best prediction result experimented.

K. USE OF MACHINE LEARNING TO PREDICT THE ONSET OF DIABETES [18]

Vinaytosh Mishra, Dr. Cherian Samuel, Prof. S.K.Sharma3 published a paper to predict diabetes using machine learning. They used Logistic Regression to predict diabetes. In their data, they used Age, Smoking, Parental Diabetes Mellitus, Hypertension & Waist Circumference, Sex, BMI and HBA1C information as the attribute. The data analysis was conducted using the software tool IBM SPSS 20.0. In result, they found the likelihood 78.5565%, Cox & Snell R Square Nagelkerke Square .628, and Nagelkerke R Square 0.839.

L. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES [19]

This paper described the aspect of different data mining algorithms like Decision Tree and Naïve Bayes Classifier in determining diabetes in women. These classification algorithms of data mining have been applied to the Pima Indians Diabetes(PID) Database of National Institute of Diabetes and Digestive and Kidney Diseases. Naïve Bayes, have been used to create the model for diagnosis. The dataset was divided into a training set and test set by the cross-validation technique and percentage split technique. 10-fold cross-validation was used to prepare training and test data. After data pre-processing in CSV format, the J48 algorithm was employed on the dataset using WEKA tool, after which data were divided into "tested positive" or "tested-negative" categories depending on the final result of the decision tree.

M. NOVEL APPROACH TO PREDICTION OF DIABETES USING CLASSIFICATION MINING ALGORITHM [3]

This work explored the performance of Decision Tree as Data Mining techniques in determining diabetes. The major aim was to forecast whether the patient has been affected by diabetes or not using the data mining tools and the MV dataset. This dataset contains 1024 complete instances of 26 Parameters. MV dataset was collected from various district's people using Questionnaires. They experimented Decision Trees to predict diabetes for local and systemic treatment. The effectiveness of the proposed model was explained with the result of a proper experimental analysis. This study discovered that the parameters of Workplace, Age, Occupation and Job Satisfaction are one of the reasons for diabetes.

N. LOGISTIC REGRESSION AND SVM BASED DIABETES PREDICTION SYSTEM [20]

This paper developed a system to predict diabetes at an early stage. To predict diabetes via three different supervised machine learning methods including SVM, Logistic regression was the main objective of this work. The clinical raw data was extracted for data pre-processing module. Then using SVM based processing and Logistic regression based processing the data was analyzed for diabetes risk prediction.

O. DETECTION OF DIABETES USING GENETIC PROGRAMMING [21]

This paper used Genetic Programming(GP) for diabetes prediction. They obtained the result with two different GP process and compared the result. They used Pima Indian Diabetes Dataset and found 78.2±2.5% accuracy for standard Genetic Programming(GP) and 78.4% ±2.2 for Comparative Partner Selection(CPS).

III. BACKGROUND STUDY

A. Data Mining

To generate new information from large sets of data can be described as data mining. Generally, data mining can be thought as the extraction of new data but instead, it is all about extrapolating patterns and new knowledge from the data that is already collected. The process may involve different technique starting from classification, clustering, association, outlier detection, regression, prediction etc. The goal of data mining is to find a consistent pattern and systematic relationship between variables to predict unknown values or future values. The use of data mining is used in almost every sector- Business, Healthcare, Education, Economics etc. But our research focuses on the healthcare industry to predict diabetes from collected medical data.

Healthcare sectors have a vast amount of structured and unstructured complex data such as patient history and records stored in written or in electronic devices, hospital resources, disease diagnosis etc [22]. These data's can be studied and processed for extracting knowledge which aid medically making a bettern prediction of disease and saving a lot of the cost. The data mining methods and tools are applied to find out hidden patterns which can help doctors and healthcare professionals as an extra source of knowledge for disease recognition and better decision making [23].

TABLE I
COMPARATIVE INFORMATION ABOUT RELATED WORKS

Authors	**Used Data Mining Approaches**	**Obtained Accuracy (%)**
VelidePhani and Lakshmi	Naïve Bayes , JRip J48(C4.5) , Decision Tree, KNN	55.85, 65.48, 68.58, 52.58, 50.
Chang et. al	RF,C4.5,MLP, LR,SVM	82.68, 76.56, 75.53, 74.48, 69.6
Tejas et. al	SVM,LR	79, 78
Pragati and Amit	CKNN, KNN, SVM, SVM+LDA, NB, SVM, ID3, C4.5,C5.0, CART	78.16, 71.84, 77.60, 88.10, 75.61, 68.3, 75.3, 76.3, 72.1
Asir	NB, MLP, RF,(WOPP) NB, MLP,RF(WPP)	76.53, 84.7, 78.25, 76.96, 84.93
Thirumal and Nagarajan N.	Naïve Bayes, C4.5,SVM,KNN	77.864,6 78.2552, 77.474, 77.7344
Mrs. Sonali	Naïve Bayes, Decision Tree, Proposed	64.0, 77.10, 83.20
Shetty	ID3	94
Kayaer	GRNN, RBF,MLP	80.21, 68.23, 77.18
Barakat	SVM, SQRex SVM, C5.0, CART, Eclectic, Jripper	89, 94, 95, 94, 93, 89
Pradhan, Madhavi et al.	Cross Validation Technique	73.0469
Aiswarya	Naïve Bayes, 48 Classification Algorithm	79.5652, 76.9565

B. Knowledge Discovery in Databases (KDD)

is a process where the low-level data is converted into high-level knowledge. It involves selecting a dataset and focusing on a subset of variables or data sample, on which knowledge discovery is to be performed. The process consists of multiple states starting from Data selection, Data cleaning, Data transformation, searching pattern, finding presentation, interpretation and finding evaluation.

Data selection - data relevant to the analysis task are retrieved from the database. From many various sources, the target data's are selected and integrated. The selection aims to focus on the correct subset of data samples and variables. The selected target dataset will be analyzed as data in the real world is incomplete, noisy and inconsistent.

Data Pre-processing is a very important step in the data mining process. Data directly taken from many heterogeneous sources will likely have inconsistencies, errors, missing values. These can be reduced through Preprocessing using data reduction techniques. If the analyzed data are not screened for such problems then it can result in misleading output. Data cleaning during Preprocessing fill the missing values, noisy data, outliers and solve inconsistencies.

Data Transformation transforms data from one format to another, that is more appropriate for data mining. The process involves basic operations like selection and construction of attribute, data smoothing, normalizing and generalization etc.

During **interpretation or evaluation**, interpretation or evaluation, multiple data sources are combined and the models and patterns are interpreted. The output results are translated in an understandable form. The evaluation of data mining result is done with statistical validation and testing its importance. Different kind of knowledge need various representation association, clustering, and classification etc. [23]

C. Classification

It is a process that generally involve two steps, one is a learning step and another one is a classification step. In learning step a classification model is established based on the training data, whereas in a classification step the model is used for prediction class label for new data. For a given test data, the accuracy of the classifier is the percentage of test set tuple which is accurately classified by the classifier.

Classification is a category of supervised learning whereas clustering falls in the category of unsupervised learning. With both algorithms, the machine can learn and able to improve from experience. These processes seek to find patterns from a certain type of data through observation and to make a better decision based on instructions and examples. Let us discuss both methods and their related algorithms used in prediction.

1) Supervised Learning: The data mining task of training and testing a machine using labeled data is known as Supervised learning. The idea is that the machine algorithm starts analyzing the data from a well-known training dataset and then testing the labeled data by modeling a function to predict a future outcome. The advantage of using supervised learning are:

- The definition of the classes is very specific. The algorithm can be trained to differentiate classes and set a boundary over an ideal decision.
- The number of classes can be determined by our self.
- The data that is used as input is familiar and labeled.
- The resulted outcomes are very much accurate compared to unsupervised learning.

Some popular supervised learning algorithms are- Naïve Bayes, Decision tree, Random Forest, Logistic Regression, K-NN, Support vector machine (SVM), Artificial Neural Network (ANN) etc. In the next sections, the algorithms are discussed elaborately.

Naïve Bayes

Naive Bayes is a well-known probabilistic classifier and it uses Bayes theorem to classify objects. The algorithm assumes the features and variables provided are independent to one another. It is carried out by using a probabilistic approach which determines class probabilities and predicts most probable classes. Naïve Bayes algorithm has been applied in a various real-world problem. It is used in real-time prediction, multiclass prediction, text classification, spam filtering, sentimental analysis, Recommended system, Digit recognition, weather prediction, categorizing news, face recognition, medical diagnosis etc.

The use of Naïve Bayes has its own pros and cons. Its advantages are that, it is very simple, fast and easy to implement. It needs less training data and highly scalable where it scales linearly with the number of predictors and data points. It able to make a probabilistic prediction, handle continuous and discrete data, and can be used for binary and multi-classification problems. On the other hand, its disadvantages are – the strong independence assumption result a loss in accuracy and practically the variables have the dependency on one another. It is considered one of the most effective machine learning algorithms. In a comparison experiment done in 2006 by *Rich Caruana* & *Alexandru Niculescu-Mizil*, naïve Bayes performed better than any other classifier algorithm [24]. In another performance study done in 2015 by *Jyoti Soni et.al,* naïve Bayes outclasses decision tree and K-NN method with an accuracy of 52.33 percent and it took 609ms to provide the output [25]. The algorithm works on the Naïve Bayes formula known as Bayes theorem that is given below.

$$P\left(class_c|X\right) = \frac{P\left(X|class_c\right) * P\left(class_c\right)}{P\left(X\right)}$$

Fig. 1. Bayes Theorem Formula.

- P(class$_c$|X) is the posterior probability of class (class$_c$) given predictor X.
- P(class$_c$) is the prior probability of class.
- P(X|class$_c$) is the likelihood which is the probability of predictor given class.
- P(X) is the prior probability of predictor.

Naïve Bayes is a popular algorithm for its use in disease prediction due to its simplicity and fast result.

Decision Tree

A Decision tree is one of the important classifiers as it is easy and its implementation is simple. Using a decision tree, a dataset is broken down into smaller and smaller subsets while at the same time an connected decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision trees are quick to build and easy to interpret. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. Furthermore, the algorithm used in the decision tree are ID3, C4.5, C5, J48, CART and CHAID which are used to measure information gain or ratio. But their efficiency is well determined over small data.

Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm, it is successfully applied to various pattern recognition problems and also a training algorithm for learning classification and regression rules from data. SVM separates the classes of data through a hyperplane and also separating class members from non-members in the input space [27]. The system automatically recognizes a fragment of informative points called support vectors and uses them to represent the separating hyperplane. Its parameter is regularized to avoid over-fitting. There are two types of SVM classifier which are Linear SVM Classifier where the dataset is linearly separable and Non-linear SVM classifier where datasets are not linearly separable. SVM is powerful over vast attributes dataset.

Random Forest

The random forest consists of many trees, and it makes a prediction by averaging the predictions of each component tree. In every iteration of the algorithm it constructs a random decision tree and after a large number of decisions tree is created, the outputs are aggregated and the result represents a strong ensemble.

The Random forest has three key advantage – firstly random forest can be used for both classification and regression, secondly, overfitting make the result worse, but if there are enough trees in the forest, the classifier model can avoid overfitting. Thirdly, Random Forest can handle missing values, and it can be modeled for categorical values.

Logistic Regression

Logistic regression is a statistical method to analyze a dataset in which there is more than one independent variable to predict an outcome. Its goal is to use the best fitting model to illustrate preferred dichotomous variable over a set of independent variables. The logistic function can be represented by the given formula

Here the formula represents the logistic function where ‘e’ is the base of natural logarithms and ’z’ is the actual numerical value that we want to transform.

$$g(z) = \frac{1}{1 + e^{-z}}$$

Fig. 2. Logistic Function

K-Nearest Neighbors

K nearest neighbors is a non-parametric technique lazy learning algorithm used in various data analysis implementation such as pattern recognition, database, mining of data due to its peak accuracy and simplicity. Non-paramedic means the structure of the model decided from the dataset and it is called lazy learning because it does not need any training data for model creation, it is done in the testing phase. It is counted under top 10 algorithms in data mining [3].

In KNN the letter ‘ K ’ represents its nearest neighbors. The main deterministic factor is of KNN is its neighbors. To predict a label of an unknown point P its closest neighbors K points need to find and classify the point P by most votes given by its neighbors [28]. Every individual object vote for their class and the class with the most votes are taken as a prediction. Euclidean distance, Hamming distance, Manhattan distance are used to find the distance between points e.g A and B in which the Euclidean distance function is the most widely used one. For points A and B are represented by feature vectors $A = (x_1, x_2, ..., x_n)$ and $B = (y_1, y_2, ..., y_n)$ where n is the dimensionality of the feature space. To calculate the distance between A and B, the normalized Euclidean metric is generally used by

$$dist(A, B) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2)}{n}}$$

Artificial Neural Network (ANN)

It is a system inspired from the brain, constructed for the purpose to replicate the way human learns. It consists of input and output layer and a hidden layer , which consist of units that transform input layer into something that the output layer can use [29]. ANN works very well in finding patterns that are very complex or difficult for a programmer to extract and teach the machine to understand and recognize. A diagram below shows the layers of ANN

2) Unsupervised Learning: In unsupervised learning the training data is unknown. The data is not labeled which means the input variables are given with no corresponding output variable. In unsupervised learning, the algorithms are left to themselves to find unknown pattern and structure. It is useful for less complexity and real-time prediction. K means clustering and hierarchical clustering are some of the popular unsupervised learning examples.

K Means Clustering

K means clustering follow a simple procedure, its goal is to find groups or cluster represent by the variable K. The algorithm works by assigning each data point to a certain cluster based on their features that are given. Each cluster is characterized by its centroid, or center point to labeled new data and they should be placed as far as away from one another to get a better result.

Hierarchical Clustering

Hierarchical clustering is also a process of grouping data, concurrently over a variety of scales of distance, by creating a cluster tree. The tree is not a single set of clusters, as in K-Means, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows determining at what scale or level of clustering is most applicable.

D. Evaluation and Testing

To estimate a model accuracy based on randomly partitioning a given data, the following techniques are used:

1) Cross Validation: Cross validation is a statistical data mining method used to determine the skills and performance of a machine learning model. The dataset is divided and portioned into n folds to use for training and testing. The process is repeated n times for training and testing. It is used commonly to compare and select model as it is easy to implement and altogether have a lower bias. In K-fold cross validation the K represents the number of groups that a given dataset is to be split into [30]. For a specific value of k chosen, if k=10 then it becomes 10-fold cross validation. So, for 10-fold cross validation, the data is divided into 10 parts in a way that each part is about the same size to one another. The process then can be repeated 10 times, with each of the subsamples used only once as the validation data. The results of 10 folds then can be averaged to produce a single estimation.

2) Confusion Matrix: It is a technique used to describe the performance of a classification model and to determine how well a classifier identify tuples of different classes. It is used to expose the connection between outcomes and predicted class. The above classifier table can be used to predict disease in the way such that if it is:

- True Positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- True Negatives (TN): We predicted no, and they don't have the disease.
- False Positives (FP): We predicted yes, but they don't actually have the disease.
- False Negatives (FN): We predicted no, but they actually do have the disease

The accuracy of a model can be found by using formula –

$$accuracy = \frac{TP + TN}{P + N}$$

and the rate of error –

$$error\ rate = 1 - accuracy_{error\ rate}\frac{FP + FN}{P + N}$$

3) Performance Evaluation: There are two matrices used to determine the performance of data mining which is a recall and false positive rate (FPR).

For a given dataset the recall can be found by using the formula

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

and False Positive Rate -

$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

.

The precision of the dataset can be found by –

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

.

4) Holdout Method: Holdout method is a similar model like cross validation but it is much simpler. The data is randomly portioned. For instance, for a value of k=3, the two third data is trained for model construction and one third for estimating the accuracy. The estimation is less optimistic because only a small part of the data is used to obtain the model.

5) Bootstrap: Bootstrap is a technique for measuring statistics of the machine learning model. It is a very pessimistic model which works very well with small datasets. During training, each part of the data are selected individually and it is likely to be selected again.

E. The Dataset Mining Tools

To make the data mining approach more convenient and faster heterogeneous tools with interesting features are being used by the researchers to analyze the dataset and select the appropriate algorithm for the final system. he popular tools for data mining in the context of diabetes risk prediction are included in this section.

1) Weka (Waikato Environment for Knowledge Analysis): Weka is a machine learning software written in Java, developed at the university if Waikato, New Zealand. Weka is an open source software available at GNU (General Public License). It contains a collection of visualization tools and algorithm for data analysis and predictive modeling with graphical user interfaces for easy functionality access [31]. Weka inherits a collection of a machine learning algorithm for solving real-world data mining problems. It supports several standard data mining tasks especially data pre-processing, regression, clustering, classification, visualization, and feature selection. The feature or attribute available in Weka is of many types, Nominal: One of the predefined list of values, Numeric: A real or integer number, Date, String, Relational. Its key feature is that it is platform independent and open source and consists of various algorithms for data mining and machine learning.

2) RapidMiner: RapidMiner is another data mining software produced by the company of the same name which provides an integrated environment for data preparation, deep learning, machine learning, text mining, and predictive analytics. RapidMiner utilized for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, validation of model and data optimization [32]. In a study of Blood Research, it is found that RapidMiner provides almost 99 % of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. RapidMiner provides data mining and machine learning procedures including - data loading and transformation, data preprocessing visualizing, predictive analytics and statistical modeling, data evaluation and deployment.

3) R programming: R programming is a language and open source software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The language R is widely used by the statisticians and data miners for developing statistical software and data analysis [33]. R programming was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and right now it is developed by the R Development Core Team (of which Chambers is a member). R and along with its libraries implement a vast variety of statistical and graphical techniques, including linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering etc. R programming can be easily extended by functions and extensions, and the community of R is noted for its active contributions. Many of the R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

4) The Natural Language Toolkit (NLTK): NLTK is a leading platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It was developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania. It consists of text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning [34].

5) IBM SPSS Modeler: IBM SPSS Modeler is a powerful, flexible data and text analytics tools that help to build accurate predictive models quickly and intuitively, without programming. It displays patterns and trends in structured and unstructured data easily using unique visual interface supported by advanced analytics. Chosen from a complete range of advanced analytical functions, including state-of-the-art algorithms, automated data preparation and rich, interactive visualization capabilities. It has access to all of IBM SPSS Modeler's predictive capabilities, as well as IBM SPSS Statistics' data transformation, hypothesis testing, and reporting capabilities, from a single interface.

IV. Problems and Furture Work Direction

Diabetes has become a widespread fatal disease over the world where the majority is considered to be in Africa and Asia. However, in the countries of Asia like Bangladesh or India, a large number of diabetic don't even know that they are carrying this deadly disease. If a person can identify that he/she has a risk of diabetes and get to know at its early stage, then it will be easier to delay diabetes with proper medication and maintenance guidance. As the uses of the INTERNET and technology increases day by day we can provide instant help to our user through it. On the other hand, web technology has quickly become the world's most common way of searching data and services in the developing world. Thus a simple website could be undertaken to provide all information for marginal users. A website providing both predictions of the risk of diabetes at the early stage and some useful health tips for both the diabetic and non-diabetic can be a viable extension of this research.

V. Conclusion

The potentiality of diabetes is increasing among the young to old age people. The present study says that detection of diabetes at its early stage can play a leading role for treatment. Simple awareness measures such as low sugar diet, proper diet can avoid obesity. As the Data mining methods, techniques and tools are becoming more promising to predict the diabetes and also reduces the treatment cost, its role in this medical health care undeniable. The main contribution of our study is to provide a systematic overview of Data Mining, Knowledge Discovery and its popular prediction algorithms, evaluation methods and dataset analysis tools used for prediction of diabetes. A comparative review of the related works is also provided to select the most common algorithms of better accuracy. We observed that the performance of algorithms differ based on the amount of data and also on the criteria of the dataset. We found that, algorithms like SVM or Bayesian Network performed with the best accuracy in most cases, whereas the Naïve Bayes classifier algorithm is easy to implementable and unaware of the size of the dataset. Finally, some direction for future work is also provided based on the analysis.

References

[1] S. Wild, G. Roglic, A. Green, R. Sicree and H. King. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 27(5): 1047- 1053.

[2] Vrushali Balpande, Rakhi Wajgi," Review on Prediction of Diabetes using Data Mining Technique", *International Journal of Research and Scientific Innovation (IJRSI)*Volume IV, Issue IA, January 2017 | ISSN 2321–2705.

[3] The 6 Different Types of Diabetes. (March 5, 2018). The Diabetic Journey, Retrieved from https://thediabeticjourney.com/the-6-different-types-of-diabetes/

[4] R Manimaram,De.M.Vanitha, "Novel Approach to Prediction of Diabetes using Classification Mining Algorithm". *International Journal of Innovative Research in Science,Engineering and Technology*,Vol 6, Issue 7, july 2017.

[5] Lakshmi V. & Thilagavathi K. (2012). "An Approach for Prediction of Diabetic disease by using b-Colouring Technique in Clustering Analysis". *International Journal Of Applied Mathematical Research*, 1(4), 520-530.

[6] DIABETES: A NATIONAL PLAN FOR ACTION. THE IMPORTANCE OF EARLY DIABETES DETECTION, ASPE, 12/01/2004.

[7] Prakash Mahindrakar, Dr. M. Hanumanthappa, "Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges", *Int. Journal of Engineering Research and Applications*,ISSN : 2248-9622, Vol. 3, Issue 6, No-Dec 2013, pp.937-941

[8] VelidePhani Kumar, Lakshmi Valide. A data mining approach for prediction and treatment of diabetes disease. *International Journal of Science Inventions Today*, 2014;ISSN 2319-5436

[9] Chang, Huan-Cheng; Chang, Pin-Hsiang; Tseng, Sung-Chin; Chang, ChiChang; Lu, Yen- Chiao, - "A comparative analysis of data mining techniques for prediction of postprandial blood glucose: A cohort study". *International Journal of Management, Economics and Social Sciences (IJMESS)*, ISSN 2304-1366,

[10] Pragati Agrawal, Amit kumar Dewangan - "A BRIEF SURVEY ON THE TECHNIQUES USED FOR THE DIAGNOSIS OF DIABETES-MELLITUS". *International Research Journal of Engineering and Technology (IRJET)*. e-ISSN: 2395 -0056; p-ISSN: 2395-0072.Volume: 02 Issue: 03 | June-2015 .

[11] Tejas N. Joshi, Prof. Pramila M. Chawan - "Diabetes Prediction Using Machine Learning Techniques". S. Dewangan.et.al. *Int. Journal of Engineering Research and Application* ,ISSN: 2248-9622, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[12] Muhammad Akmal Sapon, Khadijah Ismail and Suehazlyn Zainudin - "Prediction of Diabetes by using Artificial Neural Network". 2011 *International Conference on Circuits, System and Simulation* IPCSIT vol.7 (2011) © (2011) IACSIT Press, Singapore.

[13] Asir A.G Singh, E.J. Leavline and B. S. Baig, "Diabetes Prediction Using Medical Data", *Journal of Computational Intelligence in Bioinformatics*, vol. 10, no.1 (2017) pp.1-8.

[14] Ahmed (2016a) - "Using Data Mining To Develop Model For Classifying Diabetic Patient Control Level Based On Historical Medical Records".

[15] Ahmed (2016b). "Developing a Predicted Model for Diabetes Type 2 Treatment Plans by Using Data Mining".

[16] Rabina1, Er. Anshu Chopra2 - "DIABETES PREDICTION BY SUPERVISED AND UNSUPERVISED LEARNING WITH FEATURE SELECTION". ISSN: 2454-132 ,(Volume2, Issue5)

[17] Priya B. Patel, Parth P. Shah, Himanshu D. Patel - "Analyze Data Mining Algorithms For Prediction Of Diabetes". 2017 IJEDR | Volume 5, Issue 3 | ISSN: 2321-9939

[18] Vinaytosh Mishra, Dr. Cherian Samuel, Prof. S.K.Sharma3 - "USE OF MACHINE LEARNING TO PREDICT THE ONSET OF DIABETES" . *International Journal of Recent advances in Mechanical Engineering (IJMECH)* Vol.4, No.2, May 2015

[19] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes using Classification Mine Techniques", *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.5, No.1, January 2015.

[20] T. N. Joshi, P. M. Chawan, "Logistic Regression and SVM Based Diabetes Prediction System", *International Journal for Technological Research In Engineering*, vol. 5, issue. 11 (2018)

[21] Muhammad Waqar Aslam and Asoke Kumar Nandi - "DETECTION OF DIABETES USING GENETIC PROGRAMMING". *18th European Signal Processing Conference (EUSIPCO-2010).*Aalborg, Denmark, August 23-27, 2010

[22] Thirumal P.C and Nagarajan N, "Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus-A Case Study", *ARPN Journal of Engineering and Applied Sciences*, Vol 10 No.1 January 2015

[23] Prakash Mahindrakar, Dr. M. Hanumanthappa, "Data Mining In Healthcare: A Survey of Techniques and Algorithms with its Limitation and Challenges", *International Journal ofEngineering Research and Applications*, ISSn: 2248-9622, Vol.3 Issue 6, Nov-Dec 2013.

[24] Abid Sarwar, Vinod Sharma, "Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2", Special Issue of International Journal of Computer Applications (0975 – 8887) on Issues and Challenges in Networking, *Intelligence and Computing Technologies* – ICNICT 2012, November 2012.

[25] Ajinkar Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "A Review on Prediction of Multiple Disease and Performance Analysis using Data Mining and Visualization Techniques", *International Journal of Computer Applications (0975-8887)* Volume 155-No 1. December 2016

[26] Pragati Agrawal, Amit Kumar Dewangan, "A Brief Survey On The Techniques Used For The Diagnosis of Diabetes - Mellitus", *International Research Journal of Engineering and Technology (IRJET)* Volume: 02 Issue: 03 June 2015

[27] Vrushali Y Kulkarni," Random Forest Classifiers : A Survey and Future Research Directions", *International Journal of Advanced Computing*, ISSN:2051-0845, Vol.36, Issue.1, April 2013
bibitemi Durga Kinge, S.K. Gaikwad, "Survey On Data Mining Techniques For Disease Prediction", *International Research Journal of Engineering and Technology (IRJET)* Volume: 05 Issue:01 Jan 2018

[28] Sadegh Bafandeh Imandoust, Mohammad Bolandraftar,"Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", *S B Imandoust et al. Int. Journal of Engineering Research and Applications*,Vol. 3, Issue 5, Sep-Oct 2013, pp.605-61.

[29] Ms. Sonali. B. Maind, Ms. Priyanka Wankar, "Research Paper on Basic of Artificial Neural Network", *International Journal on Recent and Innovation Trends in Computing and Communication*,Volume: 2 Issue: 1.

[30] N.A Diamantidisa, D.Kalisb, E.A Giakoumakisa, "Unsupervised Stratification of Cross-Validation for Accuracy Estimation, *Artificial Intelligence 116(2000)*, 1-16, 25 Sep 1998

[31] Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA"*International Conference in "Emerging Trends in Science, Technology and Management*-2013

[32] Jovica Krstevski, Dragan Mihajlov, Ivan Chorbev, "Student Data Analysis with RapidMiner", *ICT Innovations 2011 Web Proceedings* ISSN 1857-7288.

[33] Anshul Jatain, Amit Ranja, "A Review Study on Big Data Analysis Using R Studio", *International Journal of Computer Science and Mobile Computing*, Vol.6 Issue.6, June- 2017, pg. 8-13.

[34] Anubha Sharma, Ritu Patidar, Rupali Dave, "Study of Different Data Mining Tools-An Overview",*International Journals of Advanced Research in Computer Science and Software Engineering* ISSN: 2277-128X (Volume-7, Issue-6).

[35] Shetty, SR Priyanka, and Sujata Joshi. "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique." IJ Information Technology and Computer Science (2016).

[36] Kayaer, Kamer, and Tulay Yıldırım. "Medical diagnosis on Pima Indian diabetes using general regression neural networks." *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*. 2003.

[37] Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." *IEEE transactions on information technology in biomedicine* 14.4 (2010): 1114-1120.

[38] Pradhan, Madhavi, et al. "Design of classifier for detection of diabetes using neural network and fuzzy k-nearest neighbor algorithm." *International Journal of Computational Engineering Research* 2.5 (2012): 1384-1387.