PREDICTION AND CLASSIFICATION USING RANDOM SUBSPACE CONDITIONAL PROBABILITIES TECHNIQUE FOR HEALTHCARE DATASETS

S. N. Santhalakshmi
Ph.D Research Scholar (Part-time), Department of Computer Science,
Nandha Arts and Science College, Erode, Tamil Nadu, India.
{snsanthalakshmi@gmail.com}
Dr. S. Prasath
Assistant Professor & Research Supervisor, School of Computer Science,
VET Institute of Arts and Science (Co-Edu) College, Erode, Tamil Nadu, India.
{softprasaths@gmail.com}

Abstract — Today there is increase in society suffering from Diabetes disease and this number is rising continuously. Diabetes is a chronic disease that leads to numerous amount of death each year. Untreated diabetes troubles the proper functionality of other organs in mankind. Hence, identifying diabetes is very important to save the human life. Data mining is the process of analyzing data based on different factors and summarizing it into useful information. Prediction is one of the mostly used techniques in medical data mining. The main aim of this work is to discover new patterns to provide meaningful and useful information for the public. The data are collected from clinic as well as in repository. The clinical data have some unknown values. Data mining techniques are applied to healthcare datasets to explore satisfactory methods and techniques in order to extract useful patterns with high accuracy with unknown values also. Generally, decision tree classifies the data it won't predict and this paper proposes an enhanced method which boosts up and develops the traditional classification algorithm for prediction. The proposed method is evaluated in WEKA tool with proper evaluation measures to confirm its efficiency.

Keywords: Classification, Prediction, Decision tree, Random subspace, Conditional Probabilities, Random forest, MLP.

I. INTRODUCTION

Digital data is data that shows other forms of data by using specific machine language systems that interprets by a variety of programming [1]. The binary system is the most important of these systems. Which commons complex audio, video and also text detail in a series of binary characters, traditionally the ones and zeros, or the values “on” and “off.” The greatest power of digital data is that all very complex analog inputs can be expressed with the binary system. With small microprocessors and large data centers, this details capture model has helped parties such as businesses and government agencies explore new frontiers in data collection and represent more accurate models.

i. Healthcare

Data mining holds big potential for the healthcare sector to enable health systems to completely use data and analytics to identify inefficiencies and best practices that improve care and reduce costs [5]. Authority believe the opportunities to improve care and reduce costs concurrently could possibly apply to as much as 30% of overall healthcare spending. But due to the difficulty of healthcare and a slower rate of technology adoption, our sector lags behind these others in performing effective data mining and analytic strategies. Like analysis and business intelligence, the style of data mining can mean different things to different people. The most main definition of data mining is the analysis of large data sets to discover patterns and use those patterns to predict the trend of future events.

ii. Diabetes

Diabetes is a disease that forms when your blood glucose, also called blood sugar, is too high. Blood glucose is your first source of energy and gets there food you eat. Blood tests are conducted to determine the diabetes [2] by evaluating the excess body glucose in blood and them urine sugar test also conducted to determine urine sugar in level.

A. Type 1 Diabetes

It can develop at any age, but occurs most commonly in children and adolescents. If you have type 1 diabetes [4], your body produces very small or no insulin, which means that you need daily insulin injections to maintain blood glucose under control level.

B. Type 2 Diabetes

It is more common in adults and accounts for around 90 percentages of all diabetes cases. When you have type 2 diabetes, your body could not make good use of the insulin that it produces. The cornerstone of type 2 diabetes treatment is healthy lifestyle, including increased physical activity and healthy diet. However, over time most people with type 2 diabetes will require oral drugs or insulin to keep their blood glucose levels under control.

C. Gestational Diabetes

Gestational diabetes is a type of diabetes that includes of high blood glucose during pregnancy and is associated with difficulties to both mother and child. It is usually leaves after pregnancy but women affected and their children are at increased risk of developing type 2 diabetes later in life.

iii. Predictive Model- Classification

Classification models predict categorical group labels; and prediction models predict continuous valued functions. For example patient can be classified as high danger or low danger. Based on the disease pattern, classification approach is used to reveal the hidden pattern. This process predicts a group label from training data set. There are various types of classification technique used to determine the diabetes. Prediction is nothing but decision out the knowledge or some pattern from the large amounts of dataset. It is used to predict missing or unavailable numerical data values rather than group labels. Prediction in data mining is to find out data points purely on the description of another linked data value. It is not necessarily linked to future events but the used variables are hidden. Prediction derives the relationship between a thing you know and a thing you need to predict for future source.

II. RELATED WORKS

This section is to provide the general overview of related works in the field of diabetes.

Minyechil Alehegn et,al., [13] proposed in this Intelligence so that be used for prediction, recommendation and recovery from disease in early stages. Techniques used for datasets analysis are Random Forest, KNN, Naïve Bayes, and J48. The dataset from UCI repository. PIDD and 130-US hospital dataset were considered. PIDD involves 768 records and 8 characteristics with one target class and 130–US hospital dataset consists of 93743 instances and 48 features. Data pre-processing has done using integrating WEKA tool. When dataset becomes large the accuracy of the proposed algorithm is not good relatively. NB and J48 prediction algorithm are better for large datasets analysis. KNN technique is not good for large dataset analysis.

Senthil Kumar et,al., [14] proposed performance of the classification is affected due to the existence of high dimensionality in medical data. Hence novel techniques Improved Firefly (IFF) and hybrid Random forest algorithm is proposed for feature selection and classification. The PIMA dataset is utilized in our proposed approach for diabetic's prediction. Data pre-processing has done using integrating WEKA tool. That the hybrid Random forest algorithm obtain the better accuracy compared to other approaches such as SVM, NB, KNN, ANN and Random forest.

Punnee Sittidech et,al., [7] the Random Forests, ensembles of weak decision trees, can be improved by excluding less important features from the model. The objective of this paper was to create a base-line, which will be useful for the classification on diabetes complications data. We recommend using the Random Forest with Feature Selection technique for other type of classification problems. all diabetes dataset were collected from Sawan pracharak Regional Hospital,Thailand. Data pre-processing has done using integrating Matlab tool. Random Forest with Feature Selection gave the best result with Feature Selection achieved increased classification performance.

Bharathidason et,al., [8] proposed has been made to improve the performance of the model by including only uncorrelated high performing trees in a random forest. This leads to inappropriate and poor ensemble classification decision. In random forest, randomization would cause occurrence of bad trees and may include correlated trees. Dataset on the Risk factors were collected from 6073 diabetic subjects of MV Diabetics Lab., Chennai. An enhanced random forest algorithm incorporating a tree selection step based on the calculated tree importance and correlation. To improve the classification accuracy of random forest with the properties of strength and correlation.

Koteswara Chari et,al., [15] proposed predict the level of occurrence of diabetes and predict the level of occurrence of diabetes using Random Forest, a Machine Learning Algorithm. Using the patient's Electronic Health Records (EHR) we can build accurate models that predict the presence of diabetes. The data set consists of 19 variables for 403 of the 1046 topics surveyed for African Americans in a research to determine even if obesity, diabetes and other cardiovascular risk factors are prevalent in central Virginia. The data mining tool WEKA has been used. Random wildwood has outperformed than other algorithms. It proved to prophesy whether several were diabetic or not. It has been proved that the proposed algorithm can achieve accuracy.

Asir Antony Gnana Singh et,al., [12] the diabetes prediction system to diagnosis diabetes. To explore the approaches to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms and methods. The Pima Indians Diabetes Data Set is used. The data mining tool WEKA has been used. The MLP machine learning algorithm, UTD test method produces better accuracy compared to other methods without pre-processing method. The pre-processing method increases the accuracy for MLP machine learning algorithm except UTD test method.

Kawsar Ahmedet,al., [9] proposed discusses about different types of data mining classification algorithms accuracies that are widely used to extract significant knowledge from huge amounts of data. Here compared 20 classification algorithms by measuring accuracies, speed and robustness of those algorithms. The Pima Indians Diabetes Data Set is used. This only discusses about accuracies of different classification algorithms using WEKA toolkit. Only uses 20 classification algorithms for classify diabetes patient data perspective. Lastly find top 5 algorithms for 3 cases like total training data set, percentage split and 10 fold cross validation.

Rashedur et,al., [6] proposed to analyze the performance of different classification techniques for a set of large data. A fundamental review on the selected techniques is presented for introduction purpose. The Pima Indians Diabetes Data Set is used. The different classification techniques using three data mining tools named WEKA, TANAGRA and MATLAB. The best algorithm in WEKA is classifier with a high accuracy.

Zahed Soltani et,al., [10] proposed Different models of artificial neural networks have the capability to diagnose this disease with minimum error. We have used probabilistic artificial neural networks for an approach to diagnose diabetes disease type II. The Pima Indians Diabetes Data Set is used. The data mining tool has been used of MATLAB. The method achieved diagnosis accuracy in training phase and test phase. Both training and testing measures could identify the diabetes disease type 2 with a good accuracy.

Manimaran et,al., [11] proposed the use of Decision Tree algorithm for classification and predict Diabetes in patients. Classification is implemented by finding rules that classify data. There are several classification and Statistical methods. MV dataset, collected from various districts is used to predict diabetes Disease using Data Mining Classification Techniques. It contains 1024 complete instances with 26 Parameters. The data mining tools used weka tool. Medical predictions need higher accuracy levels and accuracy above 85% is good for early detection/prediction of diabetes.

III. METHODOLOGY

A. Multilayer Perceptron

A multilayer perceptron is a feed head artificial neural network that generates a set of outputs from a set of inputs. IT is characterized by various layers of input nodes connected as a directed graph between the input and output layers. It consists of at least three layers of nodes, an input layer, a hidden layer and an output layer. Its sometimes colloquially referred to as "vanilla" neural networks, especially when they have an individual hidden layer. Its uses backpropagation for training the network. It is a deep learning method. It is mostly used for solving problems that require supervised learning as well as research into computational neuroscience and parallel distributed processing. It is a powerful form of an Artificial Neural Network that is generally used for regression and can also be used for classification. Supervised learning algorithm can used for both classification and regression for any type of *N*-dimensional signal.

B. Random forest tree Algorithm

The random forest tree is a classification algorithm having many of decisions trees. It uses bagging and factor randomness when building each single tree to try to create an uncorrelated forest of trees whose prediction by group is more accurate than that of any individual tree. It is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. It is algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble approach which is better than a single decision tree because it reduces the over-fitting by balance the result.

C. Support Vector Machine Regression (SVMR)

To use SVMR, as this belongs to a few clusters in the categorization complexity, for comparison purposes, A_1 and A_2 in the case of regression and support vector machine at this stage is the real figure and additional variables are similar to the categorization glitches. One of the prevalent methods for the prediction of complex datasets is regression techniques. One of the prevalent approaches for forecasting complex datasets is regression models. In this analysis, by combining 3 common regression models and the forecast sum of COVID-19, the authors formulated a simple mean aggregated system.

Support vector machines (SVM) is a supervised learning algorithm. This algorithm is used for classification and regression problems. SVR is based on the same principles as SVM for classification i.e. to find a hyperplane in a d-dimensional space (d is the number of features) that uniquely classifies the data points. SVR uses a non-parametric technique, which means, the output from the SVR model does not depend on distributions of the dependent and independent variables.

SVR technique is basically dependent on kernel functions, which allows for the construction of a nonlinear model without changing the explanatory variables, which helps in better interpretation of the resultant model. In these algorithms, a hyperplane is found that separates the different features. The produced model by SVM does not depend on the training points that lie outside the margin but instead depends on a subset of the training data as the cost function.

Similarly, in SVR, support vectors find the closest data points and the actual function represented by them. To get closest to the actual curve if the distance between the support vectors to the regressed curve is maximum. A hyperplane is a function that classifies the points in a higher dimension or other words hyperplanes are the boundaries that help in the classification of the data points. If the margin for any hyperplane is maximum, then that hyperplane is the optimal hyperplane. The points which are closest to hyperplane are called support vector points and the distance of the vectors from the hyperplane are called the margins.

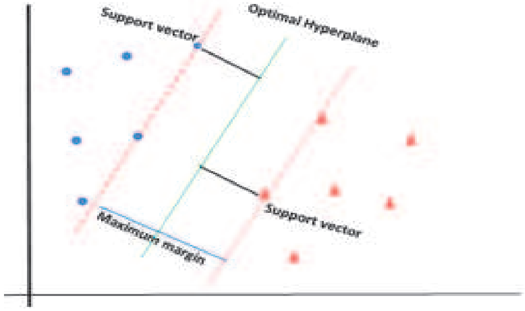


Fig.1 SVM Model Maximum-margin Hyperplane

Farther the Support Vector points, from the hyperplane, more is the probability that the points will be correctively classified in their respective region or classes. Thus, the equation of the hyperplane in the d dimension can be given as:

$$z = l_0 + l_1 x_1 + l_2 x_2 + l_3 x_3 \ \ldots. \quad (1.1)$$

$$= l_0 + \sum_{i=1}^{n} l_i x_i$$

$$= l_0 + l_1^T x$$

$$= b + l_1^T x$$

Where $l_0 = \{l_0, l_1, l_2, \ldots\ldots\}, b = biased\ term\ (l_0)\ and\ x = variables$

Kernel is an important part of SVR. The kernel is a way of computing the dot product of two vectors x and y in some high dimensional feature space. Kernel trick is used in SVR which simply means to replace the dot product of two vectors by the kernel function

D. Proposed Random Subspace combined with Conditional Probability in decision tree

This method are also as known as attribute bagging , is an ensemble learning technique that attempts to minimize the link between estimators in an ensemble by training on random model of features instead of the entire feature set. In the Random Subspace Method (RSM), one also modifies the training data. It may benefit from using random subspaces for both constructing and aggregating the classifiers. It is similar to bagging except that the features are randomly sampled, with replacement, for each learner and finally it choose the majority of voting. This method is also related to one-class classifiers. Recently, it has been used in a portfolio selection problem show its superiority to the conventional remodel portfolio essentially based on Bagging.

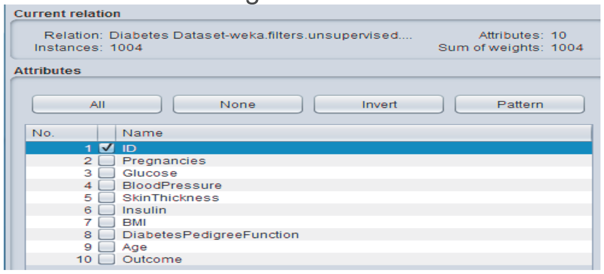
IV. RESULT AND DISCUSSION

Table 1. Shows the eight explanatory attributes and one target attribute (class) with 1004 instances taken from Kaggle dataset repository and some data are collected from nearest hospital. There are some unknown values presented in the data collected from the hospital.

Table 1. Database

Attributes	Description	
preg	Number of times pregnant since all the patients are female	
plas	Plasma	
pres	Pressure	
skin	Skin Thickness	
insu	Insulin	
mass	Body Mass Index	
pedi	Diabetes pedigree function	
age	Age of the patient	
class	Target Variable	1. Tested positive
		2. Tested Negative

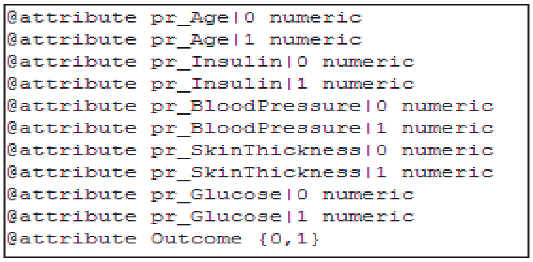
The experiment is carried out in WEKA tool which supports ARFF (attribute relation file format) file format. The file can be converted to ARFF format from Comma Separated Value (CSV) format. To give the patient a permanent identification number, in figure 2, this work assigns 'ADD-ID' method to each instances as the dataset doesn`t contains any patient name or number. Now the dataset consists of nine explanatory attribute and one target attribute.



Relation: Diabetes Dataset-weka.filters.unsupervised.... Attributes: 10
Instances: 1004 Sum of weights: 1004

No.		Name
1	☑	ID
2	□	Pregnancies
3	□	Glucose
4	□	BloodPressure
5	□	SkinThickness
6	□	Insulin
7	□	BMI
8	□	DiabetesPedigreeFunction
9	□	Age
10	□	Outcome

Fig 2. After Preprocessing



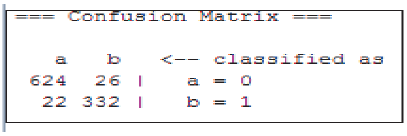
```
@attribute pr_Age|0 numeric
@attribute pr_Age|1 numeric
@attribute pr_Insulin|0 numeric
@attribute pr_Insulin|1 numeric
@attribute pr_BloodPressure|0 numeric
@attribute pr_BloodPressure|1 numeric
@attribute pr_SkinThickness|0 numeric
@attribute pr_SkinThickness|1 numeric
@attribute pr_Glucose|0 numeric
@attribute pr_Glucose|1 numeric
@attribute Outcome {0,1}
```

Fig 3. Implementation of conditional probabilities

In figure 3, Conditional probability is implemented for each attribute as per the value of class. Then this newly created dataset is feed to the decision tree.

Results –Classified Data (RSDTCP)

The dataset is classified to have patient with tested positive and negative results where (a=0) implies negative and (a=1) implies positive patients. The results are shown in confusion matrix. The diagonal element in the matrix shows the correctly classified instances whereas other elements are misclassified data.



```
=== Confusion Matrix ===

   a   b   <-- classified as
 624  26 |   a = 0
  22 332 |   b = 1
```

Fig.4 Confusion matrix

In figure 4, classified data are shown. From the correctly classified data, out of 1004 instances, 643 are tested negative and 388 are tested positive in correctly classified data. The balance (16 + 7=23) is misclassified data.

Performance analysis

Four evaluation measures are used to assess the performance of the existing and the proposed methodology. They are Accuracy, Sensitivity, Specificity and Processing time.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad --(2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad --(3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad --(4)$$

$$\text{Processing Time} \\ = \text{Time taken to train the model} \\ + \text{Time taken to test the model} \qquad --(5)$$

Where, in Equation (2-5), TN = True Negative; TP = True Positive; FP = False Positive; FN = False Negative.

Table 2. Performance analysis

Evaluation Measure	MLP	Random Forest	SVMR	RSDTCP
Accuracy	89.72	93.52	83.46	90.73
Sensitivity	88.91	93.45	83.32	90.01
Specificity	90.85	94.61	85.87	92.93
Processing time in sec	1.12	0.95	0.20	0.31

In Table 2, existing and proposed are compared and it shows the proposed method outperforms the existing in terms of accuracy, sensitivity and specificity. But the proposed takes high processing time as it calculated conditional probability and in includes Random subspace method to train the data.

V. CONCLUSION

Data mining is one of the key areas in Machine learning used to detect Diabetes disease. Though it has numerous techniques to classify, predict or group the patients, it should be enhanced to give high accuracy as this is a health sector. The enhanced algorithm should be able to handle missing values and missing labels. Hence to overcome the issue, this research proposes random subspace method combined with conditional probability in decision tree (RSDTCP). This proposed method trains the traditional classifier decision tree for prediction and checks the probability of the outcome before taking decision thus improves the accuracy. Four evaluation metrics are used to assess the performance and it shows the proposed method gives high accuracy than the existing methods. In future, the work can be extended to improve the accuracy and the work should predict the values for all genders.

REFERENCES

[1] Jiawei Han and Michelin Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, ISBN 1-55860-489-8. August 2000.

[2] *Diabetes in the UK 2010: Key statistics on diabetes* – published March 2010.

[3] Madhuri V., Joseph,” Data Mining: A Comparative study on various techniques and Methods”, Volume 3, Issue 2, Feb 2013.

[4] Anuja Kumari V, Chitra, “Classification Of Diabetes Disease Using Support Vector Machine”, International Journal of Engineering Research and Applications”, Vol. 3, Issue 2, March -April 2013.

[5] Divya Tomar, Sonali Agarwal,“A survey on Data Mining approaches for healthcare”, International Journal of Bio-Science and Bio- Technology Vol.5, No.5 , 2013.

[6] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013, 6, 85-97.

[7] Punnee Sittidech, Nongyao Nai-arun, Random Forest Analysis On Diabetes Complication Data, Proceedings of the IASTED International Conference Biomedical Engineering (BioMed 2014) June 23 - 25, 2014 Zurich, Switzerland.

[8] S. Bharathidason , C. Jothi Venkataeswaran, Ph.D, Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees, International Journal of Computer Applications (0975 – 8887) Volume 101– No.13, September 2014.

[9] Kawsar Ahmed, Tasnuba Jesmin, Comparative Analysis of Data Mining ClassificationAlgorithms in Type-2 Diabetes Prediction Data Using WEKA Approach, Internat. J. Sci. Eng., Vol. 7(2)2014:155-160, October 2014.

[10] Zahed Soltani, Ahmad Jafarian, A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.

[11] R. Manimaran and Dr. M.Vanitha, Prediction of Diabetes Disease Using Classification Data Mining Techniques, International Journal of Engineering and Technology (IJET) Vol 9 No 5 Oct-Nov 2017.

[12] Dr. D. Asir Antony Gnana Singh, Dr. E. Jebamalar Leavline, B. Shanawaz Baig, Diabetes Prediction Using Medical Data, Journal of Computational Intelligence in Bioinformatics ISSN 0973-385X Volume 10, Number 1 (2017).

[13] Minyechil Alehegn, Rahul Raghvendra Joshi, Preeti Mulay , Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48:An Ensemble Approach, International Journal Of Scientific & Technology Research Volume 8, Issue 09, September 2019.

[14] Senthil Kumar, R. Gunavathi, AN Enhanced Model for Diabetes Prediction using Improved Firefly Feature Selection and Hybrid Random Forest Algorithm, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[15] K.Koteswara Chari, M.Chinna babu, Sarangarm Kodati , Classification of Diabetes using Random Forest with Feature Selection Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019.