IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 10, October 2015.
www.ijiset.com
ISSN 2348 – 7968

Design of Hybrid Classifier for Prediction of Diabetes through Feature Relevance Analysis

N.Deepika,
PG Scholar, Department of Information Technology
Rajalakshmi Engineering College
Chennai, India
deepika.n.2014.mese@rajalakshmi.edu.in

Dr.S.Poonkuzhali,
Professor, Department of Information Technology
Rajalakshmi Engineering College
Chennai, India
poonkuzhali.s@rajalakshmi.edu.in

***Abstract*— Data mining plays a major role in the field of clinical data analysis for prediction of many critical diseases. Prediction of diabetes using data mining techniques involves several processes such as data collection, preprocessing and analysis of the collected data, Interpretation of the analyzed data, and finally the decision making process. Preprocessing is done in order to remove the inconsistencies in the collected data. The preprocessed data is analyzed and classified using different data mining classification techniques. The classifier selects the most relevant features that contribute more towards the prediction of the disease. The accuracy rate of prediction varies depending upon the type of classifier being used. Classifier evaluation is done to optimize the classifier with high accuracy. Based on the results produced by the classifier, the prediction is done. The traditional data mining techniques for the prediction of diabetes uses single classifier method for predicting the disease, which have recorded a comparatively low rate of accuracy. Thus the conceptual hybrid classifier technique is proposed to predict diabetes through Feature Relevance Analysis with high accuracy rate.**

***KEY WORDS:** Diabetes, Data Mining, Classification, Hybrid classifier, Predictive analysis, Feature Relevance analysis.*

I. INTRODUCTION

Diabetes is a life-long disease and is characterized by increased level of sugar in the blood. It is either caused due to lack of insulin in the blood or due to lack of response to insulin produced by the body. Diabetes caused by the farmer is called as Type1 diabetes and the later is called as Type2 diabetes. Type2 is the most common type of diabetes [1] and is prevalent among the adults. Factors like unhealthy diet, sedentary lifestyle, obesity and family history are the main factors contributing to type2 diabetes.

Causes for diabetes [2] may vary depending upon the Genetic behavior, Family history, and health and Environmental factors. The reason that there is no defined diabetic cause is because the cause of diabetes varies depending upon the individual and their lifestyle. The most common reason for diabetes at present is sedentary lifestyle and obesity. Sedentary lifestyle is due to the emergence of new technologies and lifestyle that have made people less active with very less physical exercise which will ultimately lead to obesity and finally result in Diabetes.

India have got over 50 million diabetic patients out of the world's 285 million. This condition will rise to more than 205 million patients all around the world by 2035 [3]. This disease is affecting more people in the working age group and is proving to be an economic burden to the country. With this huge number of diabetic patients in the country, there is no proper awareness or diabetic care programs available in our country. In order to create awareness among people about diabetes, there must be some modern technology to be followed.

One major technology growing and assisting the field of medical data analysis is Data mining. Data Mining is becoming more prevalent in the field of healthcare because there is a need for efficient methodology for detecting unknown and valuable information in Clinical data. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. Clinical data-mining involves the conceptualization, extraction, analysis, and interpretation of available clinical data for knowledge-building, clinical decision-making, and practitioner reflection.

Clinical data mining has three objectives: understanding the clinical data, assists healthcare professionals, and develop a data analysis methodology suitable for medical data.

There are numerous resources available for helping the patients improve their role in self-management of diabetes. Mobile phones and smart phones have become an integral part of daily life for many people in India. So implementation of smart phone applications and tools for managing diabetes will be an effective solution in reducing the progression of diabetes and improving the quality of life of every individual in our country. The major goal of using smart phone application is to effectively manage diabetes by improving the glycemic control and ultimately preventing from further complications of diabetes.

II. CLASSIFICATION AND PREDICTION

Classification is a supervised learning method which assigns class labels to different data groups. Classification is the most frequently used data mining function with a predominance of the implementation of Bayesian

ISSN 2348 – 7968

classifiers, Neural Networks, and Support Vector Machines. The analysis of clinical data improves the healthcare by improving the performance of patient management tasks.

The main application of data mining in the field of health services is predictive analysis. Diabetes can be predicted if the patients past history is analyzed. Predictive analysis is a data mining technique applied on electronic medical records in order to effectively predict the disease at a very early stage. After diagnosis, the patient can be medicated accordingly and monitored throughout their life.

III. SIGNIFICANCE OF THE SYSTEM

The paper mainly focuses on how classification techniques in Data mining can be applied to predict the risk factors of diabetes in the data that is being used.

The study of literature survey is presented in section IV, Methodology is explained in section V, section VI covers the experimental results of the study, section VII discuses the future study and Conclusion.

IV. LITERATURE SURVEY

Mani Butwall and Sharaddha Kumar [4] made a research to envisage the diabetic behavior in arrangement with particular life style parameters like physical activity and emotional states, based on Random Forest Classifier. They used dataset collected from National Institute of Diabetes. Digestive and Kidney diseases for the research which analyzed many classification algorithms and that Random Forest classification algorithm obtained a Forest of 100 trees with 20 splits and depth of 9 and produced the result with a very high accuracy rate of 99.7%

Eleni Georga, Vasilios Protopappas et al. [5] proposed a real time patient monitoring system called METABO system that consists of a Patient's Mobile Device (PMD), different types of biosensors, a Central Subsystem located remotely at the hospital and the Control Panel from which physicians can follow-up their patients and gain also access to the central subsystem. This Advanced system allows the physician to prescribe personalized treatment plans and frequently quantify patient's adherence to treatment. The personal monitoring device collect all the data from the patient's body and sends it to the central subsystem from which the data is analyzed by the physicians provides personalized advice to the patients.

M.M.Alotaibi, R.S.H.Istepanian et al. [6] proposed a mobile monitoring framework for monitoring the diabetic patients using smart phones. This system uses the patients mart phones in order to send the readings of the patients health condition to the central server of the hospital. The hospital in turn will analyze the patient's data and provide personalized medical advice to the patients through messages via smart phones. Though this is a very effective way to improve patient's health, this needs the patient's intervention in order to send the readings to the hospital server.

Aishwarya Iyer, S.Jeyalatha et al. [7] have done works to find a solution to the diagnosis of diabetes by applying the patterns found in the data through classification analysis by employing decision tree and naïve Bayes algorithm in WEKA where Decision tree classification has been implemented using C4.5 algorithm in order to generate the pruned tree. The confusion matrix that was obtained from these algorithms showed that the accuracy rate was about 76.95% for the farmer and 79.56% for the later.

Srideivanai Nagarajan, R.M.Chandrasekaran et al. [8] applied various classification techniques such as ID3, Naïve Bayes, C4.5 and Random Tree to improve the diagnosis of Gestational Diabetes. The system used techniques to preprocess the data and then analyzed, evaluated the effective classifier and found that the Random Tree method served the best with high level of accuracy of 0.938 and error rate of 0.062.Thus using this evaluation the pattern was obtained to predict the presence or absence of the disease.

Sanjay Kumar Sen and Sujatha Dash [9] used a combination of four supervised learning algorithms namely, Classification and Regression Tree (CART), Adaboost algorithm, Logiboost algorithm, Grading algorithm that aims at mining the relationship in diabetic data for efficient classification, compare these algorithms on the basis of misclassification and correct classification rate ,finally found that Classification and Regression Tree technique was best with an accuracy of 78.646% when compared to other meta learning algorithms.

Gunasekar Thangarasu, P.D.D.Dominic [10] proposed a framework that uses the most efficient data mining algorithms for prediction of diabetes. The framework includes neural networks, Fuzzy logic, Hybrid genetic algorithm, and Data clustering techniques. Though this framework proved to be most efficient predicting framework, this has very low tolerance to noisy data which is considered a big drawback of the framework.

Eleni I. Georga, Vasilios C. Protopappas et al. [11] proposed methodology using Support vector regression and Gaussian process for short term prediction of the disease and Clustering, classification techniques for long term prediction of diabetes. These helped in continuous monitoring of the patients and also provide personalized medical advices whenever required. Machine learning techniques were adopted to efficiently predict and monitor the patients.

V. METHODOLOGY

A. Data mining algorithms used

Classification is one of the important tasks in Data mining. There are many types of classification algorithms for classifying the data. These classification algorithms also play a significant role in analyzing and predicting the clinical data. Some of the commonly used classification algorithms for predicting diabetes are C4.5, kNN, Random Tree, Random Forest. These algorithms are used in accordance with the problem specificity. On the other hand, the algorithms have their own advantages and disadvantages. So a hybrid

ISSN 2348 – 7968

classification model is being used in this study that uses voting method for classification using J48 and IBk classification algorithm.

B. Discussion

The study consist of three main stages, data preprocessing, filtering and application of hybrid classification model viz J48 and IBk. Since data collected may not be correct always, data is preprocessed to avoid any inconsistencies in the data. Filtering is done for the feature selection process where the most relevant attributes are given highest priority while classifying the data. And finally the filtered data is classified by the hybrid data classifying system.

C. Dataset Description

The Dataset used in this work is the Pima Indian Diabetes Dataset from the UCI learning repository. This dataset is used throughout the study for classifying the healthy person and the diabetic patient [12]. This Pima Indian diabetes dataset consist of 768 instances, 8 features/attributes and one class attribute. The following table gives a brief description of the dataset.

TABLE 1. Description of the Dataset

S.No	Attribute Name	Attribute Description
1	Pregnant	Occurrence of pregnancy
2	Plasma Glucose	Concentration level noted in 2-hour OGT test
3	Diastolic BP	Diastolic Blood Pressure(in mmHg)
4	Triceps SFT	Triceps Skin Folds Thickness (in mm)
5	Serum Insulin	Serum insulin noted at 2-hour interval
6	BMI	Body Mass Index (in Kg/mm2)
7	DPF	Diabetes Pedigree Function(Family History)
8	Age	Age of person (in years)
9	Class	Has Diabetes or not

D. Data Preprocessing

Data collected is not always complete and consistent. In order to remove all the inconsistencies that are associated with the data, we go for data preprocessing. There are many data preprocessing techniques in existence. In this study, we went with the option of removing the instances that had its value as zero for it attributes Pregnant, Plasma Glucose, Diastolic Blood Pressure, Body Mass Index (BMI). In [13] it has been said that, the list wise deletion is much efficient instead of replacing the values with techniques like median, random input or mean. In our study, for data preprocessing we used fisher filtering method to derive good interval of data.

E. System Design

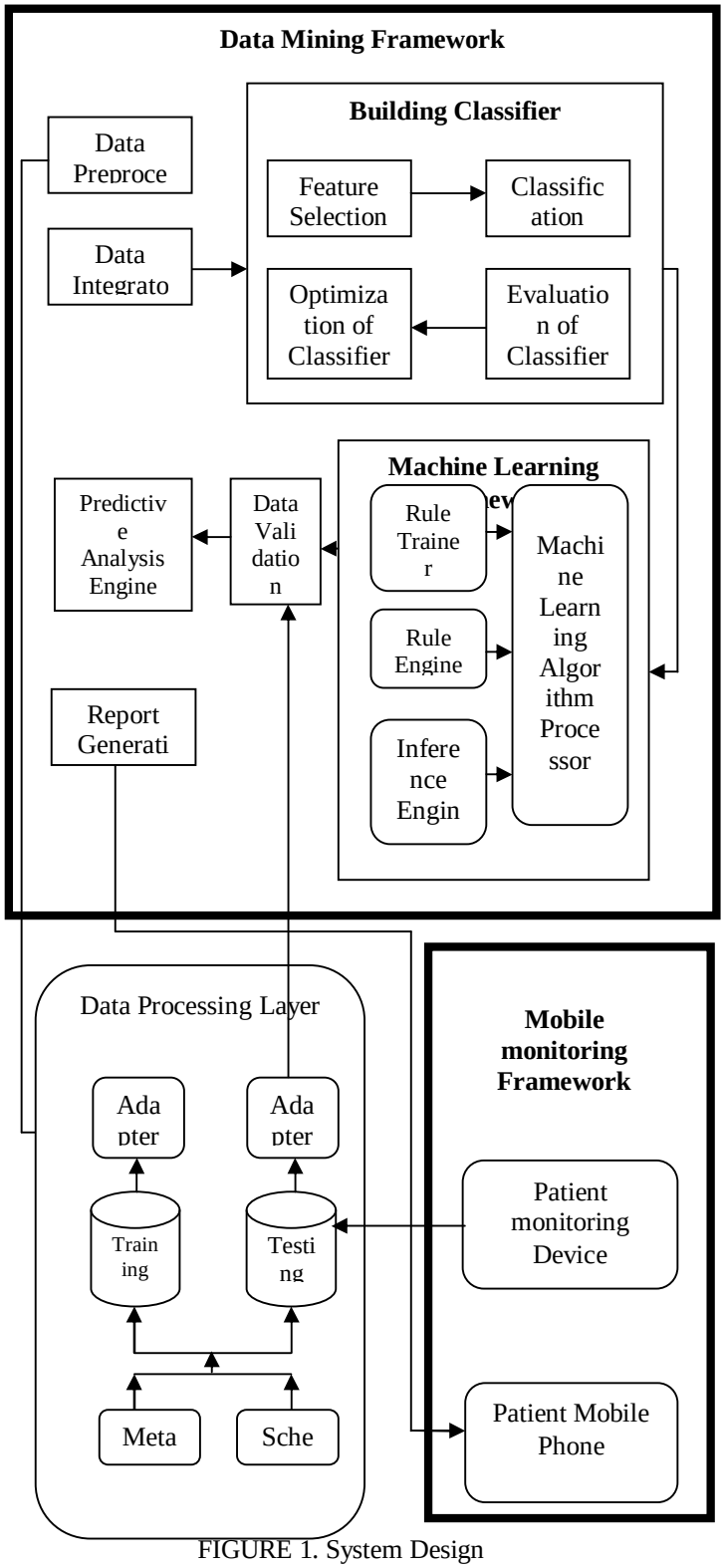


FIGURE 1. System Design

ISSN 2348 – 7968

This section explains the steps involved in building the hybrid classifier model. The system consist of four main frameworks namely the Data mining Framework, Classifier Framework, Machine learning Framework, Monitoring Framework.

The process starts with Data preprocessing followed by classifier evaluation. The machine learning framework will train the classifier with the algorithm for generating the result. Then the data is validated and finally report is being generated. The mobile monitoring framework is used for continuous monitoring of critical diabetic patients.

VI. EXPERIMENTAL RSULTS

A. *Accuracy Measure of individual classifiers*

In order to perform data analysis and prediction, a lot of data mining classification algorithms are applied in the clinical dataset and is implemented using WEKA. In this study, we have compared the performance of most of the classification algorithms such as C4.5, ID3, Random Forest, K-NN, Support Vector Machine, etc and their accuracy measures for the algorithms without filtering, Using Fisher Filter, and Correspondence analysis were noted. The Table 2 shows the accuracy measure of various data mining classification algorithms.

TABLE 2. Accuracy Measure of Classifiers

S.no	Algorithm	Accuracy percentage		
		Without Filtering	Fisher filtering	Correspondence analysis
1	C 4.5	90.62	87.7	88.15
2	MC4	79.42	79.42	79.42
3	ID3	77.21	77.21	77.21
4	K-NN	80.33	81.90	81.90
5	LDA	78.38	77.21	77.21
6	MLP	78.77	78.77	78.9
7	NBC	75.39	75.78	75.78
8	PLS-LDA	77.21	75.52	75.52
9	RND	100	100	100
10	SVM	77.47	77.08	77.08

Correspondance analysis involves finding coordinate values which represents the row and the column categories in an optimal way. The global Independence between two variables is generally measured by Chi-squared ($\chi 2$) and is calculated as,

Where E_{ij} is expected to count under independence

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij}-E_{ij})^2}{E_{ij}}$$

$$E_{ij} = \frac{N_{i.}N_{.j}}{N_{..}}$$

The figure shown below is a graph showing the accuracy rate of various classification algorithms.

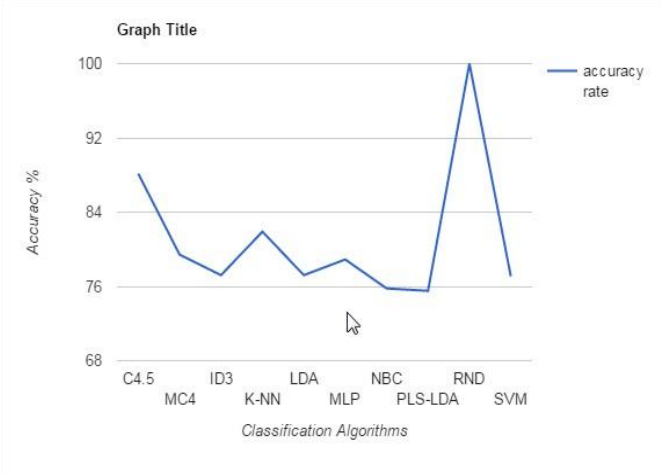
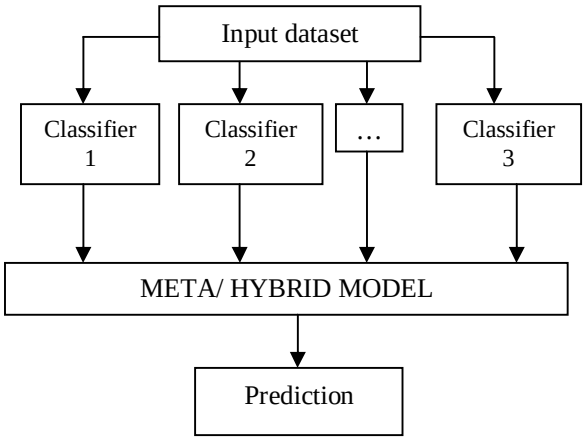


FIGURE 2. Accuracy rate of Classification Algorithms.

B. *Hybrid Classifier Model*

A hybrid classification algorithm is built using Ensemble Learning technique. There are many ensemble learning techniques. In this model we have analyzed the performance of two main ensemble learning method namely stacking and voting. The results of analysis clearly showed that voting outperformed stacking method. The general flow of the Stacking method is shown in the following figure.



ISSN 2348 – 7968

FIGURE 3. Hybrid Model

The implementation of voting method involves two major steps.

- A number of individual classifiers are trained to learn the algorithm with the training dataset. These form the first level learning classifiers.
- The individual learning classifiers are then combined by the second level learning classifiers which are called the meta-learners or the hybrid models.

The algorithm involves the following steps,

Input:

Dataset D= {(x (1) , y (1)), . . . ,(x (N) , y (N))}
First-level learning algorithms $L_1, \ldots, L_T$
Second-level learning algorithm L

Process:

For t = 1, . . . , T
h_t = Lt(D) %Train first-level individual learner h_t
End
D' = ∅ %Generate a new data set For i = 1, . . . , N
For t = 1, . . . , T
$Z_{it} = h_t(x(i))$
End
D' = D' ∪ {((zi1, . . . , ziT), y (i))}
End
h'= L(D') %Train the second-level learner h'

Output:

H(x) = h' (h1(x), . . . , hT (x))

The accuracy of the trained hybrid model is given it the table below.

TABLE 3. Classification Performance for Hybrid Methods

S.No	Hybrid Method	Algorithm	Accuracy %
1	Vote	J48 and IBK	100
2	Stack	J48 and IBK	65

It is seen that the hybrid model developed using the voting ensemble technique produced the maximum accuracy rate of 100% in predicting the Diabetic disease from the given data set.

VII. CONCLUSION AND FUTURE WORK

The proposed methodology aims at providing an efficient hybrid classification framework for predicting and monitoring the Diabetes disease. The main aim of this research is to identify and construct models that would assist medical practitioners in an efficient way by the way benefiting the people to attain longer life in this world.

The future work will explore a working model of integrating a smart phone system to this hybrid classification framework for continuous monitoring of the critical diabetic patients and also to include an automatic call mechanism to ambulance upon emergency situation if any arises for the patient who is being continuously monitored.

ACKNOWLEDGMENT

This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled "Efficient Prediction and Monitoring Tool for Diabetes Patients Using Data Mining and Smart Phone System" with Reference No: 8-169/RIFD/RPS/POLICY-1/2014-15.

REFERENCES

[1] Screening for type-2 diabetes: Report of World Health Organization & International Diabetes Federation meeting. http://www.who.in/diabetes/publications/en/screening-mnc03.pdf.

[2] An article on current status of Diabetes Mellitus in India by National Institute of Health US. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3920109

[3] http://www.idf.org/sites/default/files/Atlas-poster-2014_EN.pdf

[4] Mani Butwall, Shraddha Kumar, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications (0975 – 8887) Volume 120 – No.8, June 2015

[5] Eleni Georga, Vasilios Protopappas, Alejandra Guillen, Giuseppe Fico, Diego Ardigo, Maria Teresa Arredondo, Themis P. Exarchos, Demosthenes Polyzos, and Dimitrios I. Fotiadis, "Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System", 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009

[6] M.M. Alotaibi, R.S.H. Istepanian, A.Sungoor and N. Philip, "An Intelligent Mobile Diabetes Management and Educational System for Saudi Arabia: System Architecture", International Conference on Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS

[7] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "

ISSN 2348 – 7968

Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[8] Srideivanai Nagarajan, R.M.Chandrasekaran, and P.Ramasubramanian, "Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes", International Journal Research and Academic Review, ISSN: 2347-3215 Volume 2 Number 10 (October-2014) pp. 91-98.

[9] Sanjay Kumar Sen, Dr. Sujata Dash, "Application of Meta Learning Algorithms for the Prediction of Diabetes Disease", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 12, December 2014.

[10] Gunasekar Thangarasu, Dr.P.D.D.Dominic, "Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques", International Conference on Computer and Information Sciences (ICCOINS), 2014.

[11] Eleni I. Georga, Vasilios C. Protopappas, Stavroula G. Mougiakakou,Dimitrios I.Fotiadis, "Short-term vs. Long-term Analysis of Diabetes Data: Application of Machine Learning and Data Mining Techniques", International Conference on Bioinformatics and Bioengineering (BIBE), 2013.

[12] UCI Machine Learning Repository, "Pima Indians Diabetes Dataset", http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[13] Sonu Kumari, Archana Singh, "A Data mining Approach for the diagnosis of Diabetes Mellitus" , IEEE Journal Pg - 12, 2012, ISBN Number 978-1-4673- 4603-0.