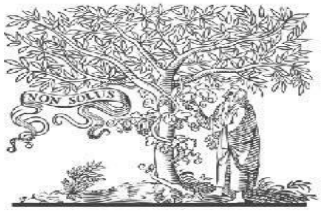
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 21st Nov 2020. Link :http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=Issue 11

10.48047/IJIEMR/V09/ISSUE 11/49

Title **AN ENSEMBLE MODEL FOR EARLY PREDICTION OF TYPICAL AND NON-TYPICAL DIABETES DISEASE**

Volume 09, ISSUE 11, Pages: 260-266

Paper Authors **P. RAVI KUMAR, D.KAMAL KALYAN, D. VIJAY KUMAR, CH. VINAY KUMAR, CH. SATISH CHANDRA, E. THARUN**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AN ENSEMBLE MODEL FOR EARLY PREDICTION OF TYPICAL AND NON-TYPICAL DIABETES DISEASE

[1]P. RAVI KUMAR, [2]D.KAMAL KALYAN, [3]D. VIJAY KUMAR,
[4]CH. VINAY KUMAR, [5]CH. SATISH CHANDRA, [6]E. THARUN
[1]Assistant Professor, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru(M), Nellore (DT), Andhra Pradesh, India
[2,3,4,5,6]B.Tech Scholars, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru(M), Nellore (DT), Andhra Pradesh, India

ABSTRACT: Diabetes among one of the most common diseases occurs in human beings due to imbalance of insulin level in blood. The early detection of diabetes is very necessary as it can affect many internal parts and immune system of human body silently. If we take proper precautions on the early stage, it is possible to take control of diabetes disease. This paper presents, An Ensemble Model for early prediction of Typical and Non-Typical Diabetes Disease. PIMA Indians Diabetes Database which is obtained from UCI repository is used as input dataset. The dataset has two types of symptoms Typical and Non-typical. An ensemble model which the combination of four classifications as SVM (Support Vector Machine), Decision Tree (DT), RF (Random Forest) and Naïve Bayes (NB). The proposed ensemble model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, and efficiently in terms of Accuracy and Precision parameters.

KEYWORDS: diabetes disease prediction, machine learning, SVM, NB, DT, RF, Typical, Non-typical.

I. INTRODUCTION

Diabetes is one of the deadliest diseases in the world. Diabetes is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin [1]. It is not solely a malady however conjointly a creator of various sorts of diseases like heart failure, blindness etc. The conventional distinguishing method is that patients ought to visit a diagnostic centre, consult their doctor, and rest for each day or additional to induce their reports.

Moreover, whenever they need to induce their diagnosing report, they need to waste their cash vainly. Diabetes Mellitus (DM) is one of the major diseases among non-communicable diseases (NCDs) which makes a huge contribution to morbidity and mortality. Moreover, DM is known as Diabetes by which a group of metabolic disorders characterized by high blood sugar levels over a prolonged period [2]. Insulin controls the level of glucose in the blood as a major hormone of the human body. At the time of the generation of insulin is diminished from islets of Langerhans in the pancreas than the Glucose level increment gradually and it causes diabetes.

Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more [3]. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels,

Blood Pressure, etc., play an important role to the contribution of the disease.

Diabetes is of 2 sorts, type 1 and type 2 diabetes. In Type 1 diabetes the beta cells of the pancreas have been hurt or attacked by the body's own protected system (auto resistance) [4]. Due to diabetic attack, A cell called beta which is responsible for generation of insulin in human body could not work properly and due to this level of insulin gets changes in human body causing high glucose (hyperglycemia). Type 1 diabetes occurs in around 5-10% which affects the person of age 30years or nearby. The signs and appearances have a quick start and are commonly genuine in nature. As Type 1 diabetes is brought about by nonattendance of insulin, people need to override what the body can't make itself.

Type 2 diabetes normally happens in grown-ups who are corpulent. There are a number of variables that subsequent to the high blood glucose levels. One of the significant factors is the body's protection from insulin in the body, basically disregarding its insulin emissions. Another factor is the falling creation of insulin by the beta cells of the pancreas. As opposed to type 2, the sort 1 diabetes most regularly happens in youngsters and is an aftereffect of the body's safe framework assaulting and decimating the beta cells. The trigger for this immune system assault isn't clear, however, the outcome is the finish of insulin creation [5].

Implementation of ML techniques can greatly improve the reach of diabetes care thus making it more effective. ML algorithms find extensive use in 4 major fields in diabetes care, such as automatic retinal screening, medical decision support, prognostic population risk stratification, and patient self-management tools. It led to a transformation in the diabetic disease management system using data-driven healthcare care. It has transformed the way diabetes is prevented, identified, and managed, which can help in reducing the worldwide prevalence of 8.8%. Various ML techniques (e.g., naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), SVM, etc.) and some deep learning techniques (e.g., Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), Deep Neural Network (DNN), Convolutional Neural Networks (CNN), deep belief network, etc.) have been constructively used in practice to detect diabetes.

II. LITERATURE SURVEY

Terry Jacob Mathew, Elizabeth Sherly et. al. [6] proposed the Analysis of Supervised Learning Techniques for Cost Effective Disease Prediction Using Non-clinical Parameters. The results show a high degree of designation accuracy for polygenic disease. This paper is use different algorithm they are Naıve Bayes gave an accuracy of 80.37% while REP trees recorded a maximum of 78.5%.Logistic regression gave 77% of accuracy. The results show a high degree of diagnosing accuracy for polygenic disease.

A.M.Rajeswari,M.Sumaiya Sidhika, M.Kalaivani C.Deisy et. al. [7] discussed and explained about “Prediction of Prediabetes using Fuzzy Logic based Association Classification”. This model is ready to predict all the categories of outliers gift in PID information set. Hence, the projected methodology is ready to work out the precise risk factors like Age, Glucose, DPF, BMI, and BP together with the proper venturous values of it to predict pre-diabetes in an improved means than the crisp

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

methodology. Mohebbi, A., Arad, T. B., Johansen, A. R., Bengtsson, H., Fraccaro, M., &Mørup, M. et.al. [8] introduced technique for recognition of type 2 diabetes utilizing profound learning in this paper the utilizations persistent glucose checking signal as a component for the recognizable proof of type 2 diabetes illness for the characterization reason they are applied convolutional neural system profound learning design and got the exactness of 77.5 % they additionally applied Logistic relapse model on the removed element where the precision was close about 65%.

Aishwarya, R., Gayathri, P., &Jaisankar, N et.al. [9] Introduced a paper on a conference where they utilize AI strategies for diabetes location here changes analyzes by utilizing bolster vector machine on the neighborhood informational index which they have been gathered from the medical clinic execution was done on Matlab 2010 and the most noteworthy exactness accomplished was 95% head segment examination and bolster vector machine combination.

Sadegh B.Imandoust, M.Bolandraftar, et al. [10] have proposed, this system that comes under the category of data mining. The system performs data mining on patterns and correlation to predict the economic events. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and also economic and financial institutions.

III. ENSEMBLE MODEL FOR EARLY PREDICTION OF DIABETES

The block diagram of An Ensemble Model for early prediction of Typical and Non-Typical Diabetes Disease is represented in below Fig. 1. The dataset aims to diagnose the patient having diabetes or not, it is based on some diagnostic measures involved in the dataset. Pima Indian dataset helps us to predict the diabetes of any Individuals with the help of our proposed methodologies. The Pima Indians Diabetes Dataset contains a total of 768 instances, with 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes. Which including no of times pregnant, glucose concentration found in oral glucose tolerance test (glucose level), blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.

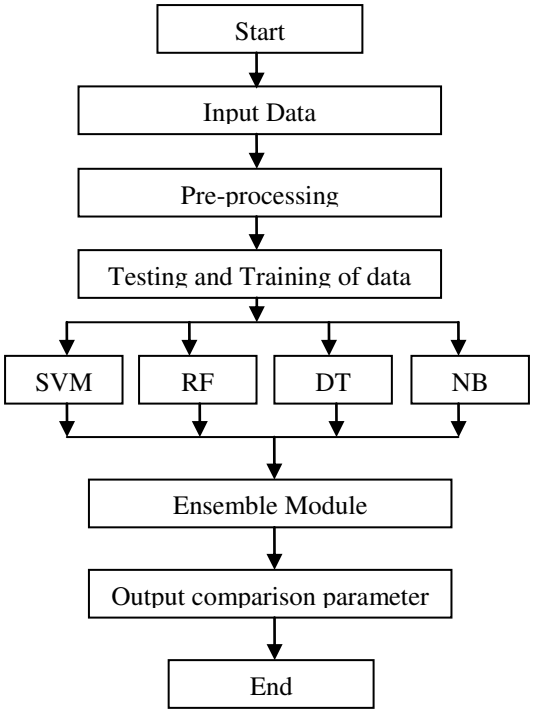


Fig. 1: BLOCK DIAGRAM OF ENSEMBLE MODEL FOR EARLY PREDICTION OF DIABETES

Preprocessing of Indian Pima Dataset can produce higher prediction accuracy. There were several missing information in the dataset. Missing values are handled with the help of calculating the standard deviation of that particular feature and allotting it to the missing spaces. So, it was essential to fill up the missing information before performing the analysis on the dataset.

Next process is data splitting. 75% of total dataset is used as training and remaining 25% of data is used for testing. After data splitting phase, the data is pass through machine learning classifiers. The combination of four classifiers is used in this paper which is called as ensemble. These four classifiers are SVM (Support Vector Machine), Decision Tree (DT), RF (Random Forest) and Naïve Bayes (NB).

A supervised algorithm for machine learning used for classification and regression purposes is the Support Vector Machine (SVM). The aim of SVM is to find the suitable margine called hyperplane between two categories. The distance from the hyperplane to the data point should be far as possible for better generalization method. The algorithms draw hyperplane that divides positive and negative dataset samples.

Random Forest is an important ensemble-supervised classification algorithm for learning which executes during the training process by building multiple decision trees. To finding prediction of value and probability estimation, Random forest method widely used. The opinion of the bulk of the trees is selected as the final choice by the random forest. Each tree in the forest treated separately according to value of random vector and equally allotted to all the trees in the forest that form a grouping of tree predictors in random forest algorithm.

Decision tree is the supervised algorithms of learning being used execute the task of classification and regression. The decision tree algorithm is also used for classification problems. It presents a structure-like tree in which each internal node will perform as a test feature of a dataset., branches act like a decision rules and every leaf represent final outcome of a tree.

Naive Bayes is a classification technique with a notion which defines all features is independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with unbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem.

After the classification of all these four type of classifiers, all results are combined in ensemble module. Then the exact output or results are evaluated by using performance metrics such as Accuracy and Precision.

IV. RESULT ANALYSIS

Pima Indian dataset helps us to predict the diabetes of any Individuals with the help of our proposed methodologies. The Pima Indians Diabetes Dataset contains a total of 768 instances, with 8 attributes. 75% of total dataset is used as training and remaining 25% of data is used for testing. Accuracy, Precision and Computational Time are used parameters used in this study for performance analysis.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Accuracy: Accuracy in classification problems is the ratio of correct predictions made by the model over all kinds of suitable predictions completed.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \ldots . (1)$$

Precision: Positive predictive value or precision is the number of accurate positive scores divided by the number of positive scores predicted by the classification algorithm.

$$Precision = \frac{TP}{(TP + FP)} \ldots . (2)$$

Where, True Positive (TP): Prediction results are yes and the patient have diabetes.
True Negative (TN): Prediction results are no and the patient do not have diabetes.
False Positive (FP): Prediction results are yes but the patient do not actually have the diabetes (Also known as a "Type 1 error").
False Negative (FN): Prediction results are no but the patient has diabetes (Also known as a "Type 2 error").

Finally we checked the computational timings of all the classifiers. In order to calculate the training and testing time, we used the python time library. The time needed for training (learning from the given data) the specific model or machine learning algorithm is known as the training time. On the other hand, the time needed for testing (checking the results by cross verifying the new data to trained data) is called as testing time. Computational time (Ct) is estimated with the following formulas below:

$$C_t = T_t + T_s \ldots \ldots (3)$$

Where, T_t , T_s represents training time and testing time, respectively.

Table 1: PERFORMANCE OF INDIVIDUAL CLASSIFIERS WITH ENSEMBLE LEARNING

Classifiers	Accuracy (%)	Precision (%)	Computational time (Sec)
SVM	82	83	150
RF	83	84	167
DT	86	82	255.7
NB	85	83	145.6
Ensemble learning (SVM+RF+DT+NB)	98	97	8.23

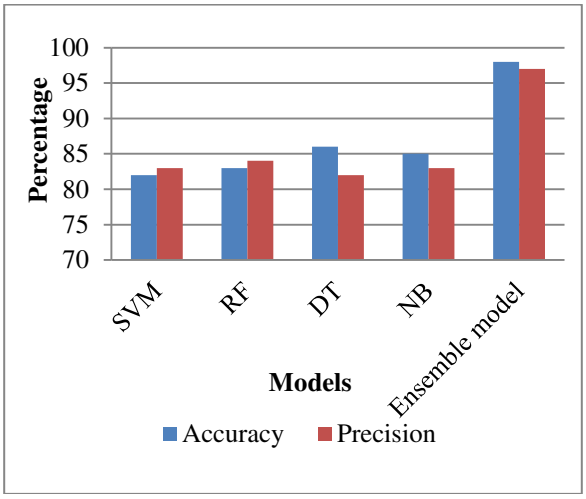


Fig. 2: COMPARATIVE ANALYSIS OF IN TERMS OF ACCURACY AND PRECISION PARAMETERS

Individual classifiers performance parameters values are compared with Ensemble learning classifier which is represented in below Table 1. Fig. 2 and Fig. 3 are shows the comparative performance analysis of Accuracy and Precision parameters, and Computational time values respectively.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

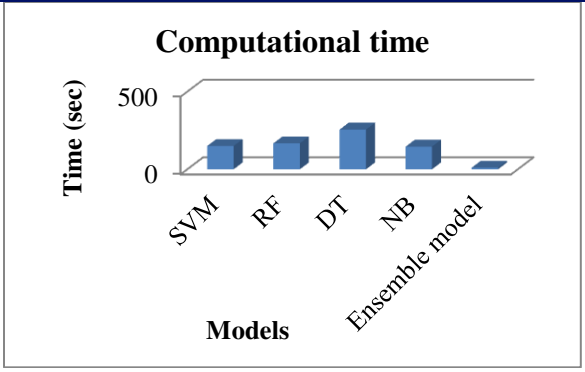


Fig. 3: COMPUTATIONAL TIME ANALYSIS

We can see that the DT classifier exerts most of the computational time and it's extremely expensive in terms of processor and memory usage and suffers from model overfitting. Ensemble classifier is efficient towards processor and memory utilization and has considerably less computational time with high Accuracy (98%) and Precision (97%).

V. CONCLUSION

In this paper, An Ensemble Model for early prediction of Typical and Non-Typical Diabetes Disease is described. The early detection of diabetes is very necessary. Experiments are performed on Pima Indians Diabetes Database. The proposed system used 768 instances in 8 attributes. 75% of total dataset is used as training and remaining 25% of data is used for testing. In order to remove unwanted data and to speed up processing time, the used data are preprocessed. The combination of four classifiers is used in this paper which is called as ensemble. These four classifiers are SVM (Support Vector Machine), Decision Tree (DT), RF (Random Forest) and Naïve Bayes (NB). Accuracy, Precision and Computational Time are used parameters used in this study for performance analysis. Ensemble classifier is efficient towards processor and memory utilization and has considerably less computational time with high Accuracy (98%) and Precision (97%).

VI. REFERENCES

[1] Hasan Abbas, Lejla Alic, Marelyn Rios, Muhammad Abdul-Ghani, Khalid Qaraqe, "Predicting Diabetes In Healthy Population through Machine Learning", 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Year: 2019

[2] Devi R. Krishnan, Gayathri P. Menakath, Anagha Radhakrishnan, Yarrangangu Himavarshini, Aparna A., Kaveri Mukundan, Rahul Krishnan , Pathinarupothi, Bithin Alangot, Sirisha Mahankali, Chakravarthy Maddipati, "Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques", 2019 IEEE Student Conference on Research and Development (SCOReD)
Year: 2019

[3] K. VijiyaKumar, B. Lavanya, I. Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes", 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Year: 2019

[4] Hamideh Fatemidokht, Marjan Kuchaki Rafsanjani, "Development of a hybrid neuro-fuzzy system as a diagnostic tool for Type 2 Diabetes Mellitus", 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Year: 2018

[5] Martina Vettoretti, Enrico Longato, Barbara Di Camillo, Andrea Facchinetti, "Importance of Recalibrating Models for Type 2 Diabetes Onset Prediction: Application of the Diabetes Population Risk Tool on the Health and Retirement Study", 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Year: 2018

[6] Terry Jacob Mathew, Elizabeth Sherly, ”Analysis Supersived Learning Techniques for Cost Effective Diaease Prediction using Non-Clinical Parameters”, sherly@iitmk.ac.in,IIITM-KTechno park, Trivndrum, July 05-07,2018.

[7] A.M.Rajeswari,M.Sumaiya Sidhika, M.Kalaivani C.Deisy,”Prediction of Pre-Diabetes using Fuzzy Logic Baesd Association Classification”, Thiagarajar College of Engineering, Madurai, India Proceedings of the (ICICCT 2018), cdcse@tce.edu

[8] Mohebbi, A., Arad, T. B., Johansen, A. R., Bengtsson, H., Fraccaro, M., &Mørup, M. (2017). “A deep learning approach to adherence detection for type 2 diabetics A Deep Learning Approach to Adherence Detection for Type 2 Diabetics”, (December).

[9] Aishwarya, R., Gayathri, P., &Jaisankar, N. (2013). A Method for Classification Using Machine Learning Technique for Diabetes, 5(3), 2903–2908

[10] Sadegh B.Imandoust, M.Bolandraftar, “Application of KNearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background”, International Journal of Engineering Research and Applications, Vol. 3, 2013