Modelling A Voting-Based Model for Diabetes Prediction Using Learning Models

DR. R. MURUGANANTHAM[1], M. SOWMYASREE[2], A.SAI RUSHIK[3], G. BHANU KIRAN[4]
[1, 2, 3, 4] *Dept. of Information Technology, TKR College of Engineering and Technology, Hyderabad, Telangana*

Abstract*— Diabetes mellitus is defined as a collection of metabolic problems that significantly impact human health worldwide. Wide-ranging study into all aspects of diabetes (diagnostic, pathophysiology, therapy, etc.) has ushered in an era of massive amounts of data. This investigation aims to provide a prediction model using machine learning, data analysis methodologies and tools in diabetic prediction. The primary goal of this work is to design a method that can more accurately predict diabetes in patients. Here, a novel ensemble model is evaluated using several characteristics such as precision, accuracy, F-measure, and recall. The machine-learning techniques are identified after hyper-tuning and cross-validation (CV) and then employed in the Vote-based ensemble model (*vEM*). According to the findings, the proposed framework can get an excellent result of approximately 92% accuracy.

Indexed Terms*— Diabetes, learning approaches, feature analysis, prediction, accuracy

I. INTRODUCTION

Diabetes is a widespread chronic condition leads to severe health risk to individuals. It is depicted by blood sugar levels more significant than usual, produced by faulty insulin sensitivity, biological effects, or both [1]. It leads to malfunction and long-term damage in diverse organs like kidneys, eyes, blood vessels, heart and neurons [2]. Type 1 diabetes (T1D) and type 2 diabetes (T2D) are the two diverse kinds of diabetes (T2D). People with type 1 diabetes are generally younger under the age of 30. Increased or frequent urination. People with type 1 diabetes are generally younger under the age of 30. Increased or frequent urination adequately with oral drugs and insulin treatment is important. Obese, hypertension, dyslipidemia, coronary artery disease, and other disorders are characteristically connected with type 2 diabetes is more widespread in older adults and middle-aged [4].

Diabetes is more widespread in people's daily lives as their living standards. As an outcome, they are learning how to evaluate and identify diabetes fast and correctly. It is predicted using glucose tolerance, fasting blood glucose and random blood glucose readings [5]. When it is predicted earlier, it will be to control. Reinforcement learning assist in preliminary prediction of diabetes mellitus based on physical gene expression profiling, and professionals can also use it as a comparison [6]. How do you choose the relevant attributes and characteristics for the machine learning approach?

Various approaches including the standard machine learning method [7] like SVM, DT and regression analysis have subsequently adopted to identify diabetes. The author in [8] used PCA and FIZ to differentiate people with diabetes from healthy persons. To diagnose type 2 diabetes, [9] adopted the QPSO method with the WLS-SVM. The scientists employed LDA to minimize the dimensionality and extraction of features in this network, the scientists employed LDA. It is a diabetes prediction method suggested by [10]. Mohapatra et al. [11] adopted a LR-based forecasting model for distinct types of type 2 diabetes onsets to handle increased datasets. Pei et al. [12] focussed on glucose and utilized support vector regression (SVR) as a regression analysis issue to predict diabetes. Furthermore, an increasing number of research employed ensemble approaches to increase accuracy [13]. Author et al. [14] introduced rotation forest, a novel aggregation strategy that integrates the learning approaches. Rashid et al. [15] suggested learning approach known as SVM prediction rules.

However, all these techniques have some flaws and drawbacks, which are effectively addressed by adopting the Vote-based ensemble model (vEM). The model gives better prediction accuracy, which is comparatively higher than other approaches.

The works are provided as follows: Section 2 analyses the anticipated Vote-based ensemble model (vEM). In section 3, the numerical outcomes attained with the proposed model are compared with other approaches, followed by the research summary in section 4.

II. METHODOLOGY

Machine learning approaches are commonly employed in diabetes prediction, producing better outcomes. Decision trees are a prominent machine learning approach with better recognition capability in the medical industry. The random forest creates a large number of decision trees. Neural networks are a relatively computational intelligence technology that outperforms conventional techniques in various ways.

a. Dataset

Diabetes dataset (2009-12) is acquired from the National Health and Nutrition Examination Survey. The dataset is an on-going research and demographic representative sample of the total population (US). There were 9858 people in the sample. If a respondent satisfied at least one of these three requirements, they were classified as a diabetic patient: plasma rising serum glucose 200 mg/dL, glucose 126 mg/dL, and glycohemoglobin 6.5%. There were around 9098 non-diabetic interviewees and 760 diabetes respondents in this study. In the dataset, there were some missing values and unexpected discoveries. There were 6561 participants in this study, including 657 diabetics and 5904 controls, after incomplete data and strange occurrences were removed from the dataset.

b. Feature learning

Using feature selection methods, you may minimize the number of characteristics and eliminate unnecessary features. There are several approaches for selecting features. We employed PCA and Minimal Redundancy and Maximal Relevance (MR^2).

c. Principal Component Analysis

By solving the algebraic equations of the correlation coefficients of the observed variables, PCA gets unit eigenvectors and K vectors. The eigenvalues, which indicate the variance of the observable variables represented by K are arranged from big to small. The following is the model for determining principal components factors:

$$F_i = T_{i1}X_1 + T_{i2}X_2 + T_{ik}X_k \ (i = 1,2,\dots,m) \quad (1)$$

Where F_i specifies i principal component factor; T_{ij} specifies i principal component factor on the j index; m specifies number of principal component factors; the number of indicators is k. PCA approach condenses many indicators. This lesser number of extensive indications capture the majority of information. They are unrelated to one another, allowing for the avoidance of questions about data. Simultaneously, reducing the number of indicators makes it easier to calculate, analyze, and evaluate data. The PCA technique was implemented using Statistical Product and Service Solutions (SPSS). It is utilized for descriptive statistics, mining, and predictive modelling.

d. Minimal Redundancy Maximal Relevance

MR^2 assures that the features possess the maximum Euclidean distances are as low as possible. The greatest applicable requirements like maximal mutual information are frequently reinforced with minimum reliability norms. The advantages are obtained in two ways. First, the MR^2 feature set have representative target for greater generalization with the same number of features. Second, we can cover the same area with smaller MR^2 features and functionality with a more significant normal feature set. The mutual method helps identify the similarity between each feature for separate explanatory data. The decision of having the most varied characteristics is known as minimum redundancies. Researchers created MR^2 for features ranking, which is similar to MR^2. They've also been used in a variety of biomedical applications.

e. Classification

Following the pre-processing of the data, the well-known ML classifiers were applied. MATLAB 2020a offers an efficient and straightforward toolbox for mining. This toolbox is utilized extensively in this

work. First, the 'train test split function from the model selection function is utilized to split the dataset into training/testing datasets. Due to the constraints resources, 90% of the dataset was utilized for training, and 10% was used for testing. Then, to diagnose diabetes, the various types of eleven classification algorithms are adapted from their relevant services.

f. Hyperparameter

The problem of selecting hyper-parameters (ML) is known as hyperparameter tuning. A hyperparameter is a value allocated to a parameter utilized to influence the learning procedure. A comparable machine-learning model may need different constraints where the learning rates to sum up varied information architectures. These parameters, known as hyperparameters, must be fine-tuned model can best address the learning problem. Hyper-Tuning will be used to acquire the best results from the ML above techniques. Cross-validation determines if numerical outputs assessing conjectured links between components are worthy of being used as knowledge representations. The K-fold cross-validation is used in this case. As a result, the dataset is split into 10 k-folds. The model selection method is used to complete this step. The k-fold CV was utilized to divide the training samples. It is used to monitor the ML classifier's CV scores, and the CV was adapted to hyper-tune. After assessing the efficiency of the classifiers, the classification techniques are recognized. The ensemble classifier is used in the ensemble's stage.

The best classifier identified before was used for these voting classifications to acquire the most remarkable performance and efficiency. Because the classifiers will reduce the complexity without significantly improving outcomes and less harm performance. The vEM is a meta-classifier that combines comparable or exceptionally good machine learning classifiers for recognition and segmentation. "Hard" and "soft" voting are carried out via the vEM classifier. The simplest example of majority voting is complex ensemble casting. The qualified majority of each classifier C_j is used to decide the class label Y:

$$Y = mode\{C1(x), C2(x), \dots, C_m(x)\} \quad j = 1,2,3, \dots, m \tag{2}$$

$$Y = argmax_i \sum_{j=1}^{m} W_j P_{ij} \tag{3}$$

where W_j is the maximum load that the jth classifier can handle.

III. ANALYSIS

To assess the classification efficacy, we employed specificity (SP), sensitivity (SN), accuracy (ACC), precision and ROC. The following are the formulas:

$$SN = \frac{TP}{TP+TN} \tag{4}$$

$$SP = \frac{TN}{TN+FP} \tag{5}$$

$$Acc = \frac{TN+TP}{TN+TP+FP+FN} \tag{6}$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \tag{7}$$

The total recognized positive samples in the affirmative set are denoted by (TP). The total categorization negative samples in the negative set is called true negative (TP). The total recognized positive samples in the negative set are the total false positives (FP). The total recognized negative classes in the positive set are represented by false negative (FN). It's frequently used to assess the accuracy of classification models. The ratio of the sample data adequately identified by the classification to the total number of instances is efficiency. There are two main qualities in medical sciences: specificity (SP) and sensitivity (SN). The actual positive rate is sensitivity, while the true negative rate is specificity. The MCC is a coefficient of correlation among actual and anticipated classifications. [-1, 1] is its value region. When the MCC equals one, the individual has made a flawless prediction. When MCC value is 0, the projected outcome isn't as superior as random forecast. When it's -1, the standard categorization is entirely different from the actual categorization. Table 1 depicts the evaluation of various prevailing approaches.

Approaches	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)	ROC (%)
k-NN	81	81	82	82	77
Adaboost	76	77	77	77	72
DT	77	77	78	77	72
SVM	75	76	75	76	73
Boosting	83	83	83	82	76
LR	76	77	77	77	72
MLP	81	81	82	81	76
NB	84	84	84	84	78
X-GB	68	70	69	69	68
GNB	80	80	81	80	74
Ensemble	92	90	91	90	93

Table 1 Comparison of proposed vs existing

This research carries out a comparative analysis for the classification of diabetic patients. In data preprocessing, to overcome the challenge of handling missing values and imbalanced class distribution problems, we have imputed the data with mean and then oversampled the data. Thenceforth, MI-based feature selection method has been utilized to obtain the high-quality data. In data classification, we have employed Tree-Based machine learning algorithms. Moreover, these algorithms are used as a base estimator of AdaBoost classifier for further improving the accuracy. It can be observed from the analysis that the Extra tree algorithm with the AdaBoost classifier outperforms other classifiers. This work has some limitations, as only MI is used for selecting features; besides, only three tree-based classifiers are applied for classification. In the future, we would like to utilize various feature selection techniques and explore different transfer learning and deep learning-based classification techniques

Random Forest is a widely used ensemble ML algorithm that germinates from Decision Trees. It engenders numerous classification models (decision trees) where each model is constructed using a feature selector such as Gini Index, Information Gain, and Gain Ratio. These models discretely learn and contribute to the prediction. The final result is made from those obtained predictions. 3) Extra Trees (ET) Extra Trees (ET) or Extremely Randomized Trees is another ensemble learning technique which combines the output of many de-correlated trees to provide its classification result. Despite having similarities, there are two fundamental differences between Extra Tree and Random Forest. First, the entire learning sample is utilized for training each tree. Second, it splits each node randomly in the learning process. Moreover, ET is superior to RF as it faster and less susceptible to noisy data. 4) Adaptive Boosting (AB) Adaptive Boosting or AdaBoost integrates many weak classifiers to generate a strong classifier. AdaBoost sets weights to each weak classifier and ensures correct classification by training the sample data in each iteration while predicting outliers or unusual observations. The intuition behind this classification technique is that a single classifier can accurately predict a portion of the dataset giving incorrect results for other portions, but incorrect portions can be correctly predicted by other weak classifiers.

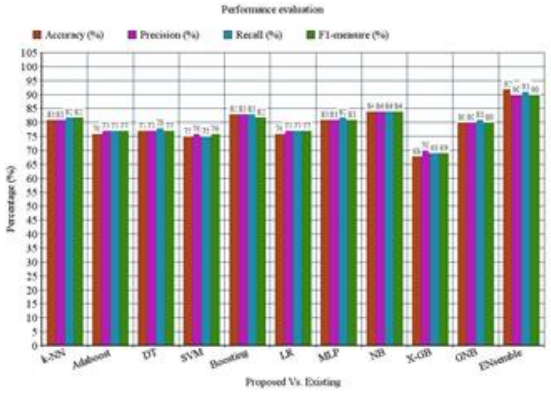

Fig 1 Performance evaluation

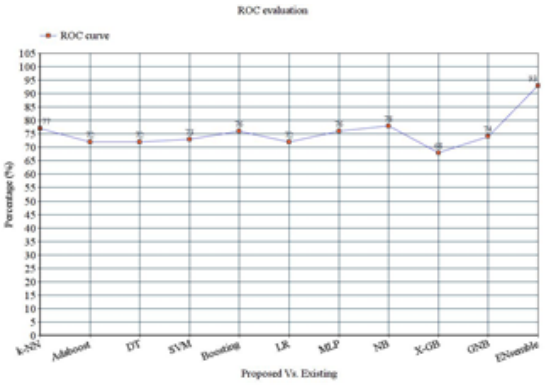

Fig 2 ROC evaluation

Table 1 depicts the comparison of diverse prevailing with the anticipated model (See Fig 1 and Fig 2). The accuracy of the anticipated model is 92% which is 11%, 16%, 15%, 17%, 9%, 16%, 11%, 8%, 24% and 12% higher than k-NN, Adaboost, DT, SVM, boosting, LR, MLP, NB, X-GB and GNB. The precision of the anticipated model is 90% which is 9%, 13%, 13%, 14%, 7%, 13%, 9%, 6%, 20% and 10% higher than other approaches. The recall of the anticipated model is 91% which is 9%, 14%, 13%, 16%, 8%, 14%, 9%, 7%, 22%, and 10% higher than other approaches. The F1-measure of the anticipated model is 90% which is 8%, 13%, 13%, 14%, 8%, 13%, 9%, 6%, 21% and 10% higher than other approaches. The ROC of the anticipated model is 93% which is 16%, 21%, 21%, 20%, 17%, 21%, 17%, 15%, 25%, and 74% higher than other approaches. Based on these analyses, it is proven that the anticipated model works well in the prediction process.

IV. CONCLUSION

Diabetes mellitus is a condition that can lead to a variety of problems. It's essential to look at how machine learning is used to forecast and diagnose this condition accurately. According to the findings, we discovered that the PCA accuracy is poor, and that the outcomes of utilizing the characteristics and MR^2 are superior. The impact of simply using overnight glucose performed better. . indicates that while fasting glucose is the essential index for predicting, we can't get the best results just by using fasting glucose. Thus we'll need another index if we want to forecast correctly. Furthermore, while comparing the effects of vEM classification, we can see that there isn't much difference between random forest, logistic regression, and neural networks. However, the random forest algorithm is superior to the other classifiers in some cases. The best moment indicates that the ensemble may be used to predict hyperglycaemia, but selecting appropriate features, classifiers, and data analysis methods is crucial. Because we can't identify the kind of diabetes based on the data, we'll try to forecast it in the future and look into the proportions of each signal to see if we can enhance the accuracy of disease prediction

REFERENCES

[1] Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya, “Diabetes in developing countries,” Journal of Diabetes, vol. 11, no. 7, pp. 522-539, Mar. 2019.

[2] Choc, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, Apr. 2018.

[3] Maniruzzaman, M. J. Rahman, M. A. M. Hasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, “Accurate diabetes risk stratification using machine learning: role of missing value and outliers,” Journal of Medical Systems, vol. 42, no. 5, pp. 92, May 2018.

[4] Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, vol. 132, pp. 1578-1585, Jan. 2018.

[5] Nai-Arun and R. Moungmai, “Comparison of classifiers for the risk of diabetes prediction,” Procedia Computer Science, vol. 69, pp. 132-142, Dec. 2015.

[6] Bansal, N. Gaur, and S. N. Singh, “Outlier Detection: Applications and techniques in data mining,” in Proc. sixth International Conference-Cloud System and Big Data Engineering, Jan. 2016, pp. 373-377.

[7] Han and H. Liu, “Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution,” Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability, vol. 23, no. 1, pp. 23-57, Feb. 2017.

[8] Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” The journal of machine learning research, vol. 15, no. 1, pp. 1929-1958, Jan. 2014.

[9] Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning

approach," Applied Computing and Informatics, Dec. 2018.

[10] Deniz E, Şengür A, Kadiroğlu Z, Guo Y, Bajaj V, Budak Ü. Transfer learning based histopathologic image classification for breast cancer detection. Health Inf Sci Syst. 2018;6(1):18.

[11] Mohapatra SK, Swain JK, Mohanty MN. Detection of diabetes using multilayer perceptron. In: International conference on intelligent computing and applications, 2019, pp. 109–116.

[12] Pei D, Zhang C, Quan Y, Guo Q. Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. J Diabetes Res. 2019; 2019:1–7.

[13] Jiang, Y., and Zhou, Z. H. (2004). Editing training data for kNN classifiers with neural network ensemble. Lect. Notes Comput. Sci. 3173, 356–361. DOI: 10.1007/ 978-3-540-28647-9_60

[14] Nabi M, Wahid A, Kumar P (2017) Performance analysis of classification algorithms in predicting diabetes. Int J Adv Res Comput Sci 8(3):456–461

[15] Rashid TA, Abdullah SM, Abdullah RM (2016) An intelligent approach for diabetes classification, prediction and description. Adv Intell Syst Comput 424:323–335