

An ensemble approach for detection of diabetes using SVM and DT

Mangalapalli Vamsikrishna¹, Manu Gupta², Jayashri Bagade³, Ratnmala Bhimanpallewar³,
Priya Shelke³, Jagadeesh Bodapati⁴, Govindu Komali⁵, Praveen Mande⁶

¹Department of Information Technology, Aditya Engineering College (A), Kakinada, India

²Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India

³Department of Information Technology, Vishwakarma Institute of Technology, Pune, India

⁴Department of Electronics and Communication Engineering, BVC College of Engineering (A), Rajamundry, India

⁵Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, India

⁶Department of Electrical Electronics and Communication Engineering, GITAM School of Technology, GITAM Deemed to be University, Visakhapatnam, India

Article Info

Article history:

Received Apr 5, 2024

Revised Nov 7, 2024

Accepted Nov 24, 2024

Keywords:

Decision tree

Diabetes prediction

Machine learning

Precision

Support vector machine

ABSTRACT

As diabetes affects the health of the entire population, it is a chronic disease that is still an important worldwide health issue. Diabetes increases the possibility of long-term complications, such as kidney failure and heart disease. If this disease is discovered early, people may live longer and in better health. In order to detect and prevent particular diseases, machine learning (ML) has become essential. An ensemble approach for detection of diabetes using support vector machine (SVM) and decision tree (DT) presents in this paper. In this case, to identify diabetes, two ML techniques are DT and SVM have been combined with an ensemble classifier. They obtain the information, they require from the Public Health Institute's statistics area. There are 270 records, or instances, in the collection. This dataset includes the following attributes: age, a body mass index (BMI) glucose, and insulin. The development of a system that predictions a patient's risk of diabetes is the goal of this analysis. Several performance metrics, including F1-score, recall, accuracy, and precision, were used to achieve this. From overall results, 96% of precision, 97% of accuracy, 96% of F1-score, and 97% of recall values are the results achieved for the ensemble model (SVM+DT) which is more effective than other individual ML models as DT and SVM.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mangalapalli Vamsikrishna

Department of Information Technology, Aditya Engineering College (A)

Surampalem, Kakinada, Andhra Pradesh, India

Email: vkmangalampalli@gmail.com

1. INTRODUCTION

Increased blood sugar or glucose levels are an early sign of diabetes, a chronic disease. A group of metabolic diseases collectively referred to as diabetes or diabetes mellitus is brought on by the pancreas inability to make insulin during metabolic activity [1]. Insulin is the hormone that the pancreas generates and it digests the glucose, which is ingested by food and passes through the bloodstream to the body's cells to provide energy. Hyperglycemia results from an increase in blood sugar levels caused by the pancreas inability to produce the insulin hormone is needed for the digestion of glucose [2]. The disease known as chronic hyperglycemia has been associated to organ and tissue failure in the body. An unnatural increase in

blood glucose levels can be caused by an insulin deficiency or by the body's cells becoming resistant to the effects of insulin. Physical weakness, itching, delayed healing, muscle stiffness, frequent urination, thirst, increased hunger, and visual blurring are among the common symptoms of diabetes [3]. It will cause a lot of problems if it is not medicated. This challenge results in death.

A healthy lifestyle and a well-balanced diet are two preventive methods that might be thought of to reduce the risk of diabetes. Regular checkups make it easier to diagnose diabetes. To identify the disease, laboratory testing are also carried out [4]. Diabetes is a metabolic disease that causes a wide range of health issues and millions of deaths worldwide every year. It is predicted that the number of diabetics in developing countries would increase from 84 million to 228 million by 2030, placing a heavy burden on all healthcare systems worldwide [5]. Therefore diabetes prediction model at early stage is the most crucial task and helps to avoid the risk of the people from the diseases that lead to cause death.

Three types of diabetes are present: type 1 diabetes, type 2 diabetes, and type 3 diabetes. When there is a shortage of insulin in the body, blood sugar levels rise and the blood sugar metabolism becomes affected. This is known as type-1 diabetes. Furthermore, obesity may result from this type of diabetes. Body mass index (BMI) increases above an individual's normal BMI are an indication of obesity [6]. Diabetes type 1 can occur throughout childhood or adolescence. The majority of food that humans consume breaks down into glucose, it is subsequently discharged into the bloodstream. Insulin, which gives energy for everyday activities, is released by the pancreas cell when blood sugar levels increase [7]. When there is not enough insulin or if the cells stop responding to the insulin, excess blood sugar remains in the blood. In the long run, this might result in major health issues like kidney disease, heart disease, and vision loss.

Adults with obesity are typically affected by type 2 diabetes. This type of condition occurs when the body either cannot make insulin or resists observing it. Type 2 typically affects middle-aged or older groups [8]. The lack of an additional provided hypoglycemic agent causes the disease to develop. "This type was referred to as "polygenic disease mellitus that is not insulin-dependent". Overweight is typically the cause.

Diabetes type 3, sometimes referred to as gestational diabetes is characterized by hyperglycemia brought on by changes in hormone levels during pregnancy [9]. In addition, there exist additional causes of diabetes, including diseases caused by bacteria or viruses, food toxins or chemicals, autoimmune reactions, obesity, unhealthy eating habits, lifestyle changes, pollution from the environment, and so on [10].

As the danger of diabetes becomes more well known, machine learning (ML) models have been applied in several recent studies as a decision-making support for early disease detection. These models allow people to start preventive measures earlier since they have an excellent level of reliability in identifying diabetes based on an individual's current condition. ML algorithms typically identify the desired approximate outcome by identifying hidden patterns within a large dataset [11]. Three categories exist for ML, a field of artificial intelligence: reinforcement learning, supervised learning, and unsupervised learning. In this system, they examine the accuracy of various common ML techniques using supervised learning algorithms. Algorithms using supervised learning attempt to predict new outcomes by using their previous understanding of the pattern observed in pre-existing data [12]. ML methods are used to identify data that is already available, such as data that is function- based, rule- based, tree- based, instance- based, or probability-based. In order to support medical specialists, different ML algorithms are introduced utilizing different data mining algorithms [13]. When compared to a single classification model, an ensemble approach in ML has demonstrated improved accuracy by combining the output from several models. As a result, the ensemble technique with support vector machine (SVM) and decision tree (DT) classification models is used in this analysis. A combination of its high dimensional data handling capacity, low computation cost, and generalized performance, the SVM is among the best supervised learning algorithms. SVM can handle numerous continuous and categorical variables provides regression classification algorithms. Generating a classification on training data and a regression model into a tree structure is the main goal of the DT algorithm. To do this, decision rules or DT are used to categorize, predict, or target variables of future or new data based on information from earlier times. It is applicable to both categorical and numerical data. A DT that is complete root nodes at each level which operate as starting points or the optimal splitting attributes for testing various attributes.

Following is the arrangement of the remaining analysis. In section 2, the literature review is compiled. Section 3 presents the described ensemble approach for detection of diabetes, In section 4, the performance evaluation results of the proposed model are examined. The paper is finally summarized and concluded in section 5, which also addresses the directions for future research.

2. LITERATURE SURVEY

Saji and Balachandran [14] focuses on examining the various multilayer perceptron (MLP) training algorithms are perform in the context of diabetes prediction. The neuroscience area of artificial neural

network research, has greatly advanced artificial intelligence. They used the Pima Indian Diabetes (PID) dataset for this investigation. In MatlabR2013, the system is implemented. The PID dataset has approximately 768 instances in it. The patient's medical history serves as the input data, and the prediction of a positive or negative test result is the target output. Based on the performance analysis, it was found that the Levenberg-Marquardt algorithm produced the best results from training over all the training algorithms.

Meng *et al.* [15] explains a study that uses common risk indicators to predict diabetes. Several categorization methods, including logistic regression, DT, and neural networks, were taken into consideration for the performance study. In terms of accuracy rate, the logistic model performed better than the other two. Family history, characteristics, and lifestyle risk are frequently included in the attributes that are studied.

Paul and Karn [16] The study evaluates diabetes detection methods based on artificial neural networks. With the use of the PID dataset on Kaggle, the study presents a study on the prediction of diabetes using k-fold cross validation and scaled conjugate gradient back propagation of artificial neural networks. 768 diabetic patients between the ages of 21 and 81 provided the data used to train the network. A hidden layer's neuronal count determines the reliable the results. During testing, the suggested technique approximates the presence and absence of diabetes represents a minimum accuracy percentage of 77% is a maximum accuracy percentage of 100% using 8 input attributes.

Zhao and Yu [17] Using the concept of model migration for online glucose prediction, a quick model building approach for new subjects is suggested. Techniques: first, a base model is created using a priori knowledge or one that may be experimentally recognized from any subject. In order for the updated models to accurately represent the unique glucose dynamics generated by inputs for new subjects, the parameters of the base model's inputs are then appropriately corrected based on a small amount of additional information from new subjects. The suggested approach can be thought of as a more efficient and cost-effective modeling approach than the difficult, subject-dependent modeling approach, particularly in cases where modeling data.

Pustozarov [18] a DT gradient boosting algorithm-based data-driven blood glucose model was developed and full detailed in order to forecast several elements of postprandial glycemic responses. The patient features, glycemic index, eating context (details of prior meals), among the meal-related information the model used from a mobile app diary were behavioral analyses. Utilizing random search cross-validation to select parameters, several models for gradient boosting were trained and evaluated. Two hours after consuming food, the most accurate models are used to measure the increased area under the blood glucose curve. Fazakis *et al.* [19] the area under the receiver operating characteristic (ROC) area under the curve (AUC) for the ensemble weighted voting logistic regression, random forest (LRRFs) ML model is 0.884, is suggested as a way to enhance diabetes prediction. With regard to the weighted voting, the best weights are determined by calculating the ML model's associated sensitivity and AUC using a bi-objective evolutionary algorithm. Additionally, a comparison between the Leicester and finish diabetes risk score (FINDRISC) systems, a number of ML models, utilizing both inductive and transductive learning, is shown. The English longitudinal study of ageing (ELSA) database provided the data used in the research.

Nuankaew *et al.* [20] suggests using factors that indicate individual health situations to predict the start of type 2 diabetes. An effective prediction model requires the individual to have many health problems resulting from different individual attributes. Based on this assumption, this paper proposes an original prediction technique known as average weighted objective distance (AWOD). The proposed methodology was validated by looking at a total of 392 entries from two widely accessible datasets: PID (dataset 1) and Mendeley data for diabetes (dataset 2). In comparison to current ML -based approaches, the suggested strategy outperformed them in terms of accuracy, according to the results, providing 93.22% and 98.95% for datasets 1 and 2, respectively.

Mahmood and Abdullah [21] analyzed the performance of five classification algorithms namely naïve bayes (NB), SVM, multi-layer perceptron artificial neural network, DT, and random forest using diabetes dataset that contains the information of 2,000 female patients. Various metrics were applied in evaluating the performance of the classifiers such as precision, AUC, accuracy, ROC curve, f-measure, and recall. Experimental results show that random forest is better than any other classifier in predicting diabetes with a 90.75% accuracy rate. Lee and Kim [22] examine the association between type 2 diabetes and the Hypertriglyceridemia waist (HW) phenotype in adult Korean. Determine the predictive ability of several phenotypes that combinations of different anthropometric parameters and triglyceride (TG) levels. Using HW and individual anthropometric data, they used binary logistic regression (LR) to investigate statistically significant differences between normal people and those with type 2 diabetes. A combination of ML algorithms, LR, and NB, were utilized to evaluate the predictive capability of different phenotypes in order to produce more reliable prediction results. Described could provide useful clinical data for the creation of clinical decision support systems for type 2 diabetes initial screening.

Kangra and Singh [23] aims to identify the most informative subset of features. Diabetes is a chronic metabolic disorder that poses significant health challenges worldwide. For the experiment, two

datasets related to diabetes were downloaded from Kaggle and the results of both (datasets) with and without feature selection using the genetic algorithm were compared. The researchers can better comprehend the importance of feature selection in healthcare through this study.

Lee *et al.* [24] aims to use a mix of different indicators to predict the fasting plasma glucose (FPG) status among adult Koreans, which is used in the diagnosis of type 2 diabetes. This study involved 4870 individuals in total, of which 2955 were female and 1915 were male. They examined the FPG status predictions made by two machine-learning systems utilizing individual combined assessments based on 37 anthropometric measurements. According to research, anthropometric measure combinations were better than single measures at predicting FPG status in both males and females. They demonstrate that utilizing balanced data from high and normal FPG groups can enhance prediction decrease the model's intrinsic bias in supporting the majority class.

Le *et al.* [25] presented a ML algorithm to forecast the diabetic patients would develop the condition. This newly developed wrapper-based feature selection method optimizes the MLP reduces the number of required input characteristics by utilizing adaptive particle swam optimization (APSO), and grey wolf optimization (GWO). Suggested method's computational results demonstrate that not only can less characteristics be required greater accuracy in predictions, 97% for APGWO - MLP and 96% for GWO - MLP, can also be achieved. This work may find use in clinical settings and develop into a useful resource for physicians.

3. METHOD

Figure 1 represents the block diagram of an ensemble approach to diabetes detection using SVM and DT. When giving data to the algorithm, they perform changes on it, a process known as pre-processing. Transforming the unprocessed data into a set of understandable data, preprocessing techniques are utilized. In other words, when data is collected in an unprocessed format from multiple sources, it becomes unusable for analysis. It is essential to the performance of the model.

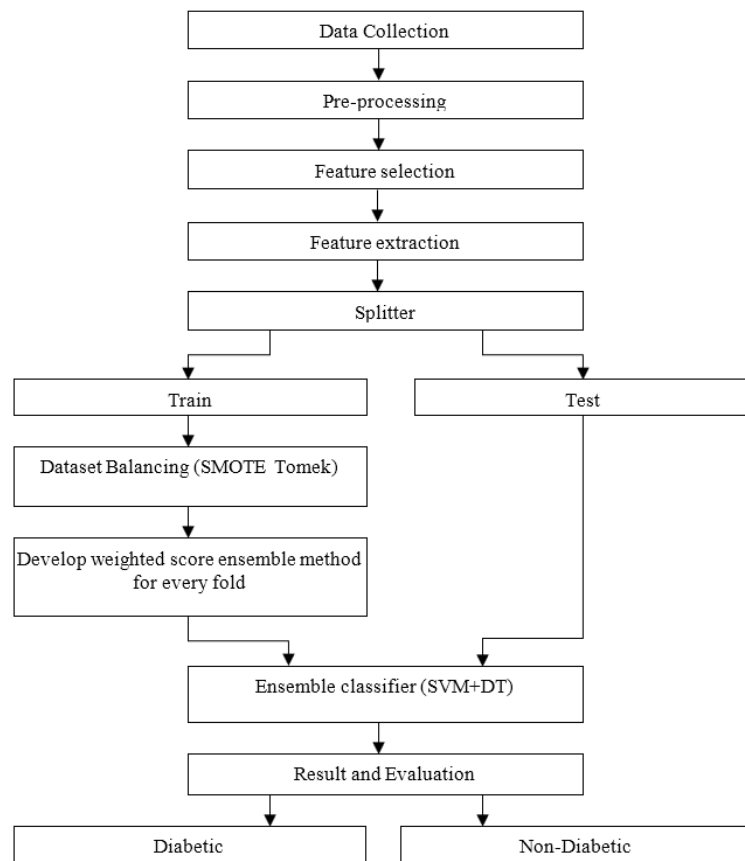


Figure 1. Block diagram of detection of diabetes using SVM and DT

Predictive modeling requires feature selection, which is done to address multicollinearity and eliminate redundant features that have strong correlations with one another in order to enhance the model's performance. In order to change the input data into the features' output, feature extraction is utilized. A characteristic of input designs that helps differentiate between the various types of input designs is attribute square measure. The input data will be considered to be redundant in the algorithm if it is too large to process. This task can be performed by using the extracted feature rather than the whole initial data set.

Eighty percent of the dataset is utilized for testing, 20% of the remaining amount is allocated to training. The ML training data set is used to teach the model to perform a large number of actions. The model is trained by retrieving certain features from the training set. To determine whether the model is exhibiting the correct actions, testing this type of data is necessary.

There are multiple ways to achieve balance in a dataset. The SMOTETomek, which combines the SMOTE and Tomek algorithms, was used in this analysis. The synthetic minority oversampling technique is referred to as SMOTE. Tomek is a method of under sampling. To get a balanced distribution of the classes, additional synthetic minority samples were first created using the SMOTE. In order to increase the separation between the two classes, the Tomek link was also utilized to eliminate samples that were near their boundaries. The test set was not changed; it was exclusively applied to the training dataset.

A weighted model was created specifically for the ensemble categorization. The two algorithms that were used in the ensemble method each received two weights. A loop was utilized to verify the set of weights that produced the best accuracy for each fold, and this combination was chosen for each fold.

To increase the model's stability and predictive ability, the separate models were integrated into an ensemble approach. When using this method instead of just one model, a better prediction performance is possible. The group develops methods for combining several ML models into a single prediction model. The ensemble model makes use of two ML algorithms: SVM and DT.

A combination of its high dimensional data handling capacity, low computation cost, and generalized performance, the SVM is among the best supervised learning algorithms. SVM can handle numerous continuous and categorical variables provides regression classification algorithms. The dimension of the classified items has not any impact on the efficiency of SVM-based classification. Using special nonlinear functions called kernels, the input space is converted into a multidimensional space, this approach provides strong discriminative power. It is evident that, for a given amount of data, selecting the appropriate kernel function and optimal parameter values is essential. Additionally, by default, all attributes are normalized.

Generating a classification on training data and a regression model into a tree structure is the main goal of the DT algorithm. To do this, decision rules or DT are used to categorize, predict, or target variables of future or new data based on information from earlier times. It is applicable to both categorical and numerical data. A DT that is complete root nodes at each level which operate as starting points or the optimal splitting attributes for testing various attributes. Branches will result from the test's yield. In order to describe or predict the new information, the leaf hub operates as the final class mark or target variable by generating connections between arrangement rules at the root and leaf.

The suggested diabetes detection model's performance is evaluated using the following performance measurements: F1-score, accuracy, precision, and recall. Suggested strategy effectively detects diabetic patients if the accuracy is high. The people with diabetes are then recommended for export for additional healthcare. If not, the patient's test sample data is evaluated to be free of diabetes.

4. RESULT ANALYSIS

This section shows the performance of described ensemble approach for detection of diabetes model. They obtain data from the Public Health Institute's statistics sector. There are 270 records, or instances, in the collection. Prior to using the technique, the dataset received some preprocessing. Eighty percent of the dataset is used for training, while twenty percent is used for testing, respectively. The efficiency of ML techniques is evaluated using many statistical evaluation measurements, including F1-score, accuracy, precision, and recall. The terminology utilized to build these categorization measurement elements are:

- False positive (FP): incorrect positive prediction.
- True positive (TP): correct positive prediction.
- False negative (FN): incorrect negative prediction.
- True negative (TN): correct negative prediction.

Accuracy: it measures the model's total number of accurate predictions and can be measured as a ratio between the number of correct prediction and total number of test cases of model as in (1).

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

Precision: the proportion of correct positive predictions to total positive predictions is known as precision as shown in (2).

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

Recall: true positive rate, sensitivity, or recall defined here is a measure that tells us what ratio of positive instances that actually have diabetes with the actual positive instances. The (3) defines the recall.

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

F1-score: is a weighted average of the recall and precision. For the good performance of the classification algorithm, it must be one and for the bad performance, it must be zero.

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Table 1 shows the comparative analysis of the performance parameters as F1-score, accuracy, precision, and recall of individual classification algorithms, such as DT and SVM, with the ensemble model (SVM+DT).

Table 1. Comparative performance analysis

Parameters	DT (%)	SVM (%)	Ensemble model (SVM+DT) (%)
Accuracy	89	92	97
Precision	88	91	96
Recall	87	91	97
F1-score	89	90	96

Figure 2 states comparative graphical representation analysis in terms accuracy for described ensemble model (SVM+DT) and individual methods as DT and SVM. It is clear that, the ensemble model's (SVM+DT) accuracy outperforms that of other individual methods. Comparative analysis of precision parameter for ensemble method and individual models is graphically represented in Figure 3 and it states ensemble model (SVM+DT) achieves higher percentage of precision compare to individual methods as DT and SVM.

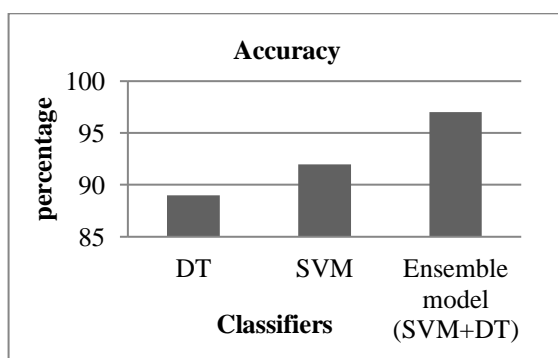


Figure 2. Comparison of accuracy

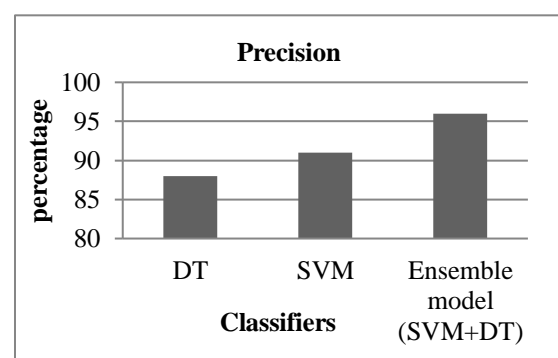


Figure 3. Comparison of precision

Figure 4 demonstrates the recall comparative analysis graphically for ensemble method (SVM+DT) and individual models as DT and SVM. It shows that ensemble model (SVM+DT) is gains higher percentage of recall than individual models as DT and SVM. F1-score based comparative graphical analysis for ensemble method (SVM+DT) and individual models as DT and SVM which is represented in Figure 5. Results states that, F1-score value is higher for ensemble model (SVM+DT) compare to individual models as DT and SVM which is efficient.

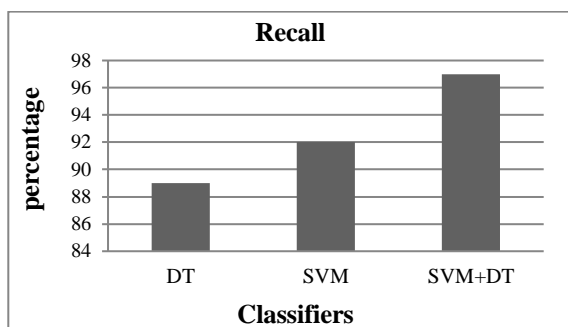


Figure 4. Comparison of recall

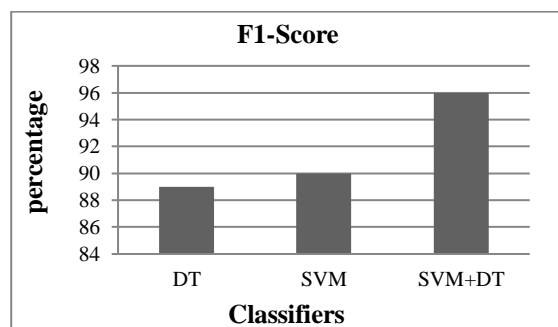


Figure 5. Comparison of F1-score

From overall results, 96% of precision, 97% of accuracy, 96% of F1-score, and 97% of recall values are the results achieved for the ensemble model (SVM+DT) that is presented. From results, it concludes that ensemble ML classification technique (SVM+DT) outperforms comparatively with individual ML models (DT and SVM).

5. CONCLUSION

An Ensemble approach for detection of diabetes using SVM and DT is described in this paper. Increased blood sugar or glucose levels are an indication of diabetes, a chronic illness. Detection of patient with diabetes at early stage is the most crucial task and helps to avoid the risk of the people from the diseases that lead to cause death. When compared to a single classification model, an ensemble approach in ML has demonstrated improved accuracy by combining the results from several models. As a result, an ensemble approach with SVM and DT classification models is used in this analysis. They obtain the information they require from the Public Health Institute's statistics sector. There are 270 records, or instances, in the collection. For the purpose of using the technique, the dataset received some preprocessing. Eighty percent of the dataset is used for training, while twenty percent is used for testing, respectively. Predictive modeling requires feature selection, which is done to address multicollinearity and eliminate redundant features that have strong correlations with one another in order to enhance the model's performance. In order to change the input data into the features' output, feature extraction is utilized. The SMOTETomek, which combines the SMOTE and Tomek algorithms, was used in this analysis. The synthetic minority oversampling technique is referred to as SMOTE. Tomek is a method of under sampling. To get a balanced distribution of the classes, additional synthetic minority samples were first created using the SMOTE. The suggested diabetes detection model's performance is evaluated using the following performance measurements: F1-score, accuracy, precision, and recall. results obtained for the ensemble model presented are 96% for precision, 97% for accuracy, 97% for recall, and 96% for F1-score. From results, conclude that ensemble ML classification technique high outperforms comparatively individual ML models. In future, we intend to implement this study to an integrated diabetes decision support system (DDSS) which is very helpful to diabetes patients.





REFERENCES

- [1] I. Aljamaan and I. Al-Naib, "Prediction of blood glucose level using nonlinear system identification approach," in *IEEE Access*, vol. 10, pp. 1936-1945, 2022, doi: 10.1109/ACCESS.2021.3139578
- [2] L. Meneghetti, A. Facchinetti, and S. D. Favero, "Model-based detection and classification of insulin pump faults and missed meal announcements in artificial pancreas systems for type 1 diabetes therapy," in *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 1, pp. 170-180, Jan. 2021, doi: 10.1109/TBME.2020.3004270.
- [3] H. Roopa and T. Asha, "A linear model based on principal component analysis for disease prediction," in *IEEE Access*, vol. 7, pp. 105314-105318, 2019, doi: 10.1109/ACCESS.2019.2931956.
- [4] W. Wang, M. Tong, and M. Yu, "Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization," in *IEEE Access*, vol. 8, pp. 217908-217916, 2020, doi: 10.1109/ACCESS.2020.3041355.
- [5] A. Ramesh, C. K. Subbaraya, and R. K. G. Krishnegowda, "A remote health monitoring framework for heart disease and diabetes prediction using advanced artificial intelligence model", *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol 30, no 2, May 2023, doi.org/10.11591/ijeecs.v30.i2.pp846-859
- [6] C. Obasi, I. Ndu, and O. Iloanusi, "A framework for internet of things-based body mass index estimation and obesity prediction", *2020 International Conference on e-Health and Bioengineering (EHB)*, 2020, doi: 10.1109/EHB50910.2020.9280202.
- [7] S. Langarica, M. Rodriguez-Fernandez, F. J. Doyle III and F. Núñez, "A probabilistic approach to blood glucose prediction in type 1 diabetes under meal uncertainties," in *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 5054-5065, Oct. 2023, doi: 10.1109/JBHI.2023.3309302.




- [8] R. N. Patil, S. Rawandale, N. Rawandale, U. Rawandale, S. Patil, "An efficient stacking based NSGA-II approach for predicting type 2 diabetes", *International Journal of Electrical and Computer Engineering (IJECE)*, vol 13, no 1, 2023, doi: 10.11591/ijece.v13i1.pp1015-1023.
- [9] D. A. K. Wardani, S. Sugiarto, and R. Cilmiaty, "Stress, nutritional status and blood glucose levels among patients with diabetes mellitus type 2," *International Journal of Public Health Science (IJPHS)*, vol. 7, no. 4, pp. 283-288, doi: 10.11591/ijphs.v7i4.14914.
- [10] P. S. Muller and M. Nirmala, "Diagnosis of gestational diabetes mellitus using radial basis function," *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, India, 2016, pp. 1-4, doi: 10.1109/GET.2016.7916859.
- [11] R. Sofiana and Sutikno, "Optimization of backpropagation for early detection of diabetes mellitus", *International Journal of Electrical and Computer Engineering (IJECE)*, vol 8, no 5, October 2018, doi: 10.11591/ijece.v8i5.pp3232-3237.
- [12] L. Flores, R. M. Hernandez, L. H. Macatangay, S. M. G. Garcia, and J. R. Melo, "Comparative analysis in the prediction of early-stage diabetes using multiple machine learning techniques", *The Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol 32, no 2, November 2023, doi: 10.11591/ijeecs.v32.i2.pp887-899.
- [13] S. Reshmi, S. K. Biswas, A. N. Boruah, D. M. Thounaojam and B. Purkayastha, "Diabetes prediction using machine learning analytics," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, 2022, pp. 108-112, doi: 10.1109/COM-IT-CON54601.2022.9850922.
- [14] S. A. Saji and K. Balachandran, "Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction", *2015 International Conference on Advances in Computer Engineering and Applications*, 2015, pp. 201-206, doi: 10.1109/ICACEA.2015.7164695.
- [15] X.-H. Meng, Y.-X. Huang, D.-P. Raon, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", *Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93-99, 2013, doi: 10.1016/j.kjms.2012.08.016.
- [16] B. Paul and B. Karn, "Diabetes mellitus prediction using hybrid artificial neural network," *2021 IEEE Bombay Section Signature Conference (IBSSC)*, Gwalior, India, 2021, pp. 1-5, doi: 10.1109/IBSSC53889.2021.9673397.
- [17] C. Zhao and C. Yu, "Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type 1 diabetes," in *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 5, pp. 1333-1344, May 2015, doi: 10.1109/TBME.2014.2387293.
- [18] E. A. Pustozarov, "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," in *IEEE Access*, vol. 8, pp. 21908-219321, 2020, doi: 10.1109/ACCESS.2020.3042483.
- [19] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," in *IEEE Access*, vol. 9, pp. 103737-103757, 2021, doi: 10.1109/ACCESS.2021.3098691.
- [20] P. Nuankaew, S. Chaising and P. Temdee, "Average weighted objective distance-based method for type 2 diabetes prediction," in *IEEE Access*, vol. 9, pp. 137015-137028, 2021, doi: 10.1109/ACCESS.2021.3117269.
- [21] I. N. Mahmood and H. S. Abdullah, "Analyzing the behavior of different classification algorithms in diabetes prediction", *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, Issue. 1, pp. 201-206, 2024, doi: 10.11591/ijai.v13.i1.pp201-206.
- [22] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," in *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 39-46, Jan. 2016, doi: 10.1109/JBHI.2015.2396520.
- [23] K. Kangra and J. Singh, "A genetic algorithm-based feature selection approach for diabetes prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, Issue. 2, pp. 1489-1498, 2024, doi: 10.11591/ijai.v13.i2.pp1489-1498.
- [24] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," in *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 555-561, March 2014, doi: 10.1109/JBHI.2013.2264509.
- [25] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic," *IEEE Access*, vol. 9, pp. 7869-7884, 2021, doi: 10.1109/ACCESS.2020.3047942.

BIOGRAPHIES OF AUTHORS






Dr. Mangalapalli Vamsikrishna     is currently working as Professor in the Department of IT, Aditya Engineering College (A), Surampalem, Kakinada. I, have around 20 + years of teaching experience in various engineering colleges, universities. I have pursued my MCA from IGNOU, later completed my M.Tech. in Computer Science from Sam Higginbottom University of Agriculture, Technology and Sciences (formerly Allahabad Agricultural Institute Deemed University (AAIDU)) and then also completed my M.Tech. in the stream of Artificial Intelligence and Robotics (AI & R) from Andhra University. Later Completed my Doctoral Degree from Centurion University of Technology and Management (CUTM) in the area of Medical Image Processing. I have around 90 + publications in various national and International Journals. Also have around 15 publications in national and international Conferences. Later, started guiding research scholars from various specializations like cloud computing, medical image processing, data science, cyber security, and machine learning. Currently, I have successfully guided 14 scholars and have been awarded with Doctoral Degrees. I also have 6 Scholars who are currently doing research under my guidance. He can be contacted at email: vkmalampalli@gmail.com.






Manu Gupta    has completed her doctorate in 2019 from Birla Institute for Technology and Science Pilani, Hyderabad Campus, India. She is currently working at the Sreenidhi Institute of Science and Technology, Hyderabad, in the Department of Electronics and Computer Engineering. Her research focuses on machine learning, deep learning, sentiment analysis, object detection and pattern recognition. She has published more than 30 articles in journals and conferences throughout the world. She can be contacted at email: manugupta5416@gmail.com.






Jayashri Bagade    is an associate professor in Information Technology Department at Vishwakarma Institute of Information Technology, Pune (India). She has graduated from BAMU University, Aurangabad, India in 2005 with Master of Engineering in Computer Science and Engineering and Ph.D. in Computer Science and Engineering from Thapar Institute of Engineering and Technology, Patiala, Punjab (India). Being interested in education field, she opted for teaching as a profession. She has 23 years of experience in the field as on date and published 30 papers in national and international conferences. Her main areas of interest are Information Storage and Retrieval, Image Processing and Soft Computing, Machine Learning. She is reviewer of reputed journal in India and abroad. She is life member of ISTE and CSI. She can be contacted at email: jayashrihedao@rediffmail.com or jayashree.bagade@viit.ac.in.






Ratnmala Bhimanpallewar    holds a Doctor of Philosophy degree in Computer Science and Engineering from K L University, Vijayawada, India. She has received her master's (M.E. Computer Science and Engineering) degree from PICT, Savitribai Phule Pune University, Pune, India. She is working as an assistant professor in the Information Technology Department of Vishwakarma Institute of Information Technology, Kondhwa (Bk.), Pune. She has 13.5 years of working experience. Her area of interest is Databases, Machine Learning and IoT. She is a lifetime member of ISTE. She has completed one funded project under ASPIRE scheme of SPPU. She can be contacted at email: ratnmalab@gmail.com or ratnmala.bhimanpallewar@viit.ac.in.






Priya Shelke    holds a Doctor of Philosophy degree in Computer engineering from Savitribai Phule Pune University, Pune, India. She has received her master's (M.Tech. Computer Science and Engineering) degree from Visvesvaraya Technological University, Belgavi in 2009. She is working as an associate professor in Information Technology Department of Vishwakarma Institute of Information Technology, Kondhwa (bk), Pune. She is having 20 years of working experience. Her area of interest is Image processing and block chain technology. She has published over 40 papers in national and international conferences and journals. She is a life time member of ISTE. She can be contacted at email: priya.shelke@viit.ac.in.






Dr. Jagadeesh Bodapati    working as a professor in the Dept. of Electronics and Communication Engineering, BVC College of Engineering(A),Rajamundry, With more than 17 years' experience I published 16 journals and 3 patents and attended 6 international conferences, 32 FDPs, 11 workshops and organized 4 workshops, Reviewed one Ph.DThesis and more than 20 journals.And ratified by university in deferent levels i.e., assistant professor and associate professor, and one student awarded Ph.D. degree and another student pursuing Ph.D Under my guidance in jntuk university and I guided many projects to under graduate and graduate students. I have played significant role in transforming my knowledge to faculty and my students, and I am committed to continually update and enhance my skill set. He can be contacted at email: bjadageesh2020@gmail.com.



Govindu Komali    is a highly motivated assistant professor with a passion for teaching and research. Experienced in the development and implementation of innovative instructional methods. Creative assistant professor with 1 year of experience in providing an engaging and stimulating learning environment. She can be contacted at email: komaligovindu1996@gmail.com.



Praveen Mande    is currently working as an assistant professor in Department of Electrical Electronics and Communication Engineering, GITAM School of Technology at Gandhi Institute of Technology and Management, Visakhapatnam. His research interests include power system operation and control, smart grids and micro grids, electrical vehicles, power electronics and power quality improvement using FACTS devices and its applications. He can be contacted at email: pmande@gitam.edu.