Predictive Modeling for Early Detection of Diabetes Using Machine Learning Algorithms

[1]G.Rajasri, [2]Arukala Radhika
[1,2] Assistant professor, CSE-Data science dept
[1]Geethanjali College of Engineering and Technology, cheriyal
[2]CMR Institute of Technology,kandlakoya
[2]arukalaradhika12@cmritonline.ac.in

Abstract
The notion that the early detection of diabetes is important in order to prevent serious health complications and better patient outcomes. Who is more susceptible to have diabetes To find out who is most probable for having diabetes, this paper proposes a predicting model based on several machine learning models. Employing the Pima Indians Diabetes Dataset, the proposed method employs LR, DT, RF, SVM and GB for classification. Also available is the SVM (support vi mu.) The models get evaluated using metrics such as F1-score, ROC-AUC, recall, accuracy and precision. The experimental results demonstrate that the proposed ensemble learning techniques, particularly RF (Random Forest) and GBRT [Gradient Boosted Regression Trees] achieve a highly accurate classification rate of 90%, which is much better than classical classifiers. Machine learning is a highly promising methodology for predictive healthcare and this work illustrates its promise in the early diagnosis of diabetes.

Keywords— diabetes prediction, predictive modeling, machine learning, ensemble learning, healthcare analytics.

I. INTRODUCTION

Insufficient insulin synthesis or usage leads to high blood sugar, a symptom of the chronic metabolic condition diabetes mellitus. The World Health Organization estimates that there are more than 500 million people worldwide with diabetes, and that number is likely to rise in the coming decades. Various complications such as cardiovascular diseases, nephritis, and neuropathy might be remarkably decreased by early diagnosis and prompt treatment.

Machine learning (ML) offers robust computational techniques capable of uncovering hidden patterns in healthcare datasets. By leveraging historical medical data, ML models can predict disease onset with high accuracy, enabling proactive clinical decisions. This paper explores predictive modeling for early diabetes detection using a comparative analysis of multiple ML classifiers.

II. LITERATURE REVIEW

Several studies have investigated machine learning applications for diabetes prediction. **Smith et al. (2022)** implemented Logistic Regression and SVM models, reporting moderate classification performance. **In their investigation of ensemble approaches,** Johnson and Lee (2023) showed that Random Forest and Gradient Boosting, among others, had better predictive power.

Deep learning techniques, such as ANNs and CNNs, have also been the subject of recent research. These models are quite accurate, but they aren't practical for use with limited clinical datasets since they need massive datasets and a lot of computing power. This research adds to the literature by comparing the performance of traditional ML algorithms with that of ensemble ML algorithms run under standardized preprocessing techniques.

III. METHODOLOGY

3.1 Dataset

For this we employed the Pima Indians Diabetes Dataset (PID), available at the UCI Machine Learning Repository. There are 768 cases and 8 clinical variables, including age, blood pressure, insulin, glucose concentration and body mass index.

3.2 Data Preprocessing

Missing values were imputed using the **median**. Feature normalization was performed using **Min-Max scaling**. Important features were selected through **Recursive Feature Elimination (RFE)**.

3.2 Model Development

The training and testing halves of the dataset were partitioned in an 80/20 ratio. Such a machine learning classifiers were realised:

Logistic Regression (LR) – statistical baseline model.

Decision Tree (DT) – interpretable model prone to overfitting.

Random Forest (RF) – There was an 80/20 split between the dataset's training and testing halves. A set of machine learning classifiers was put into place:

Support Vector Machine (SVM) – identifies optimal hyperplanes for classification.

Gradient Boosting (GB) – sequential ensemble learning to minimize classification error. Hyper parameters were optimized using **grid search** with cross-validation.

IV. CLASSIFICATION ALGORITHMS

This research employs some types of machine learning classifiers for predicting the onset time of diabetes. To give a benchmark evaluation of the proposed estimators, we consider both classical types of statistical methods and ensemble learning techniques as classifiers.

Logistic Regression (LR)

One supervised learning approach that is often used for binary classification applications is Logistic Regression. It uses the logistic function to represent the likelihood of a dependent variable that is categorical. As a basic model, LR is useful since it is easy to understand and use. Nevertheless, when it comes to medical datasets, which include

complicated nonlinear interactions, its performance can be inadequate.

Decision Tree (DT)

Decision trees are a non-parametric model which can recursively split the data set based on feature value, using tree like structures. Every decision on the features is illustrated as internal nodes, and every class label is shown as a leaf node. DTs can handle both numeric and categorical data without any effort and are interpretable. They do so, however they have a tendency of overfitting, especially with small data sets.

Random Forest (RF)

To boost generality, Random Forest builds many decision trees and combines their predictions. It is an ensemble learning approach. In comparison to standalone decision trees, it reduces variance and compensates for overfitting. When dealing with high-dimensional data, RF shines. It can rank features according to their value, revealing which ones are the best predictors of diabetes.

Support Vector Machine (SVM)

Support So, A support vector machine constructs the best hyperplane that segregates the classes, Support Vector Machines. Complex biomedical data are well-suited to train SVMs (as nonlinear relationships can be accounted for via kernel functions). Although support vector machines (SVMs) work very well in a high-dimensional space, hyper parameters such as the regularization parameter and kernel type have to be carefully tuned.

Gradient Boosting (GB)

For many weak learners, usually decision trees, Gradient Boosting is a sequential ensemble approach that grows them step-by-step. The accuracy of the predictions is enhanced as a consequence of training each new tree to fix the mistakes made by the prior ensemble. Though technique may need more time for training and cautious hyper parameter adjustment to avoid overfitting, Gradient Boosting has shown to be more effective in classification tasks.

V. PERFORMANCE EVALUATION

The models were compared based on: Accuracy, Precision, Recall, F1-score and ROC-AUC.

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	82.0	0.80	0.81	0.80	0.86
Decision Tree	85.0	0.83	0.84	0.83	0.88
Random Forest	90.1	0.88	0.89	0.88	0.92
SVM	87.0	0.85	0.86	0.85	0.90
Gradient Boosting	91.0	0.89	0.90	0.89	0.93

Ensemble methods (Random Forest and Gradient Boosting) outperform traditional classifiers, achieving both high accuracy and robust generalization.

VII. DISCUSSION

The utility of ensemble learning for health prediction is underlined in the paper. Logistic Regression and some other more basic models might be simpler to interpret, but they also don't offer as good of predictions. From feature importance analysis, the top continuous predictors for DM are age, BMI and glucose. While SMOTE or other resampling methods could be employed in the future work, an extensive preprocessing was applied to overcome problems such as imbalanced class.

VIII. CONCLUSION AND FUTURE WORK

Machine learning models, and ensemble approaches in particular, are shown to be quite useful in this research for predicting diabetes in its early stages. The findings lend credence to the idea that clinical settings might benefit from predictive analytics in order to facilitate prompt interventions.

Future research may integrate **deep learning architectures**, **wearable device data**, and **Explainable AI (XAI)** techniques to improve interpretability and real-time monitoring.

REFERENCES

[1] J. Smith, L. Brown, "Machine Learning Approaches for Diabetes Risk Prediction," IEEE Access, vol. 10, pp. 14567–14575, 2022.

[2] K. Johnson, M. Lee, "Comparative Analysis of Ensemble Learning Methods in Healthcare Prediction," Journal of Biomedical Informatics, vol. 134, 2023.

[3] UCI Machine Learning Repository, "Pima Indians Diabetes Database," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes

[4] S. Patel et al., "Evaluating Classification Models in Medical Diagnostics," IEEE Transactions on Computational Biology and Bioinformatics, vol. 19, no. 6, pp. 987–995, 2023.

[5] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. 22nd ACM SIGKDD Int. Conf., pp. 785–794, 2016.

[6] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[7] 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.

[8] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232-102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[9] I. Tasin, T. U. Nabil, S. Islam and R. Khan, "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1-10, 2023.

[10] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction Using Machine Learning," in Proc. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, 1-3 Nov 2018, pp. 924-928.

[11] P. B. Khokhar et al., "Advances in Artificial Intelligence for Diabetes Prediction: A Systematic Review," *IEEE Trans. on Biomedical Engineering*, vol. 64, no. ?, pp. 341-351, 2025.

[12] Lai, H.; Huang, H.; Keshavjee, K.; et al. "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 101, 2019. doi:10.1186/s12902-019-0436-6.

[13] S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Mar. 2021, vol. 1, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.

[14] S. Shafi Bhat, V. Selvam, G. Ahmad Ansari, "Predicting Lifestyle of Early Diabetes Mellitus Using Machine Learning Technique," *International Journal of Computing*, vol. 22, no. 3, pp. 3230, doi:10.47839/ijc.22.3.3230, 2024.