iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84743

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction: An Ensemble and Explainability-Oriented Approach

Dr. Ashish Sharma[1], Dr. Mohammad Islam[2], Dr. Avnish Bora[3]
[1]*Professor, Department of Technology, JIET UNIVERSE, Jodhpur 342802, Rajasthan, India*
[2] *Associate Professor, Department of Computer Science, Maulana Azad University, Jodhpur-342802 Rajasthan, India*
[3]*Professor, Department of Electronics and communication engineering, JIET UNIVERSE, Jodhpur 342802, Rajasthan, India*

Abstract: ***Early detection of diabetes mellitus, a chronic condition affecting a large share of the world's population, plays a major role in keeping long-term complications and healthcare costs down. Here, eight commonly used supervised classifiers — namely Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Naïve Bayes, AdaBoost, and Gradient Boosting — are put through a head-to-head comparison for the diabetes-prediction task. Training and evaluation were carried out on 768 patient records from the Kaggle-hosted Pima Indians Diabetes Dataset, spanning eight clinical attributes (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age). Accuracy, precision, recall, F1 score, AUC-ROC, and 10-fold cross-validation together formed the basis for judging each model. Among the eight, Gradient Boosting came out on top with 75.32% accuracy and an AUC-ROC of 0.8257, narrowly ahead of AdaBoost and Random Forest, both at 74.68% accuracy and 0.8253 AUC-ROC. When Random Forest was used to rank feature importance, Glucose, BMI, and Age surfaced as the strongest predictors of diabetes onset. Overall, boosting-based ensembles appear well suited to combining strong accuracy with clinical interpretability, pointing toward practical, low-cost, real-time screening tools.***
Keywords: Machine Learning; Diabetes Prediction; Classification Algorithms; Random Forest; Feature Importance; Healthcare Analytics; Ensemble Learning.

I. INTRODUCTION

At its core, diabetes mellitus is a metabolic disorder in which blood glucose stays abnormally elevated, either because the body does not produce enough insulin or because it cannot use the insulin it does produce. Worldwide, hundreds of millions of adults already live with the condition, and projections point to further growth over the next few decades [1] — a trend that will keep adding pressure to healthcare systems, especially in developing regions.

Catching the disease early matters a great deal, since it can head off serious downstream complications: cardiovascular disease, kidney damage, retinal damage, and nerve damage among them. The trouble is that conventional diagnostic methods, though reliable, are often costly, resource-heavy, and reliant on laboratory infrastructure that rural or under-resourced areas simply may not have.

Against this backdrop, machine learning has drawn growing interest as a way to build predictive healthcare tools that work directly with clinical data already being collected, flagging patients who may be at risk before symptoms become severe. By learning from historical patient records, these algorithms open the door to earlier intervention and smarter allocation of limited healthcare resources. A large share of this research relies on the Pima Indians Diabetes Dataset [2], chosen mainly because it is freely accessible and its features are well understood.

Building on that motivation, this work sets up a reproducible, side-by-side comparison of eight supervised classifiers for diabetes prediction, all run through an identical preprocessing and validation pipeline, and pairs this with feature-importance analysis to keep interpretability in view — something that counts for a lot once ML models move beyond accuracy benchmarks into real clinical decision-making. Two goals guide the work: first, gauging how well the most common ML classifiers handle a standardized diabetes-onset prediction task; and second, pinning down which clinical parameters matter most for prediction, while weighing the trade-off between raw performance and interpretability for practical healthcare use.

II. LITERATURE REVIEW

A great deal of prior research has tackled diabetes prediction with machine learning, most of it built around the Pima Indians Diabetes Dataset or comparable clinical repositories. Working with this dataset, Sisodia and Sisodia [3] compared Naïve Bayes, Decision Tree, and SVM, and found Naïve Bayes came out ahead on accuracy. In a broader comparison, Khanam and Foo [4] pitted seven machine learning algorithms against neural networks with varying hidden-layer setups; Logistic Regression and SVM proved competitive among the classical options, while a two-layer neural network topped out at 88.6% accuracy. Adding lifestyle-related features to baseline classifiers, Mujumdar and Vaidehi [5] were able to push accuracy higher, and in a related vein, Kopitar et al. [6] benchmarked regression- and tree-based models for early-stage type-2 diabetes risk, underscoring how much longitudinal clinical data adds to prediction quality. Association-rule mining paired with classification models was Rastogi and Bansal's [7] approach, and it too pointed to glucose and BMI as reliably strong predictors. Looking across several such comparative studies involving K-Nearest Neighbors, Naïve Bayes, Decision Tree, Logistic Regression, and SVM, reported accuracies tend to fall between 75% and 85% [8], with ensemble and instance-based methods generally edging out single decision trees on this particular dataset.

Beyond classical comparisons, another line of research has chased hybrid ensemble and deep-learning strategies to squeeze out further gains. In one such effort, Chao and Liu [9] combined Random Forest with SMOTE for class balancing [10] and SHAP for interpretability, reaching an AUC of 0.91 and pointing to glucose, BMI, and family history as the top predictors, consistent with what clinicians already understand about diabetes risk. Robustness gains from combining classifiers were the focus for Akula et al. [11], who built a supervised ensemble for Type 2 diabetes prediction, while Patro et al. [12] ran a large-scale comparison that likewise found ensembles beating out single classifiers on structured clinical data. Deep learning enters the picture through Aslan and Sabanci [13], who reframed tabular clinical data as images so convolutional networks could process it, and through Fan [14], whose hybrid CNN-LSTM-MLP ensemble blends convolutional and sequential learning for better diagnostic accuracy. Elsewhere, pre-trained CNNs, LSTM layers, and conditional GANs were combined for synthetic data augmentation [15] to work around the Pima dataset's limited size, and Farnoosh et al. [16] built DiabetesXpertNet, an attention-based CNN that holds its own on larger diabetes datasets. Taking a different route entirely, Noh and Kim [17] moved past correlation-based prediction into causal inference, pairing logistic regression and deep learning with causal-discovery tools like LiNGAM to trace physical activity and lifestyle choices as direct causal factors. Across this body of work, a clear pattern emerges: growing attention to interpretability, class-imbalance handling, and clinical relevance, well beyond raw accuracy — a direction this study also follows through its feature-importance analysis.

III. RESEARCH METHODOLOGY

A. *Dataset Description*

The Pima Indians Diabetes Dataset [2] forms the basis of this study, pulled directly from Kaggle. It holds 768 patient records, each carrying eight independent diagnostic attributes alongside a binary target (Outcome) — 1 for a diabetes-positive diagnosis, 0 otherwise — and every record belongs to a female patient aged 21 or above. A full rundown of the attributes appears in Table 1.

Table 1: Dataset Attribute Description

Attribute	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration (mg/dL)
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin-fold thickness (mm)
Insulin	2-hour serum insulin (mu U/mL)
BMI	Body Mass Index (kg/m²)
Diabetes Pedigree Function	Function scoring likelihood of diabetes based on family history
Age	Age in years
Outcome	Class variable (1: Diabetic, 0: Non-diabetic)

B. *Data Preprocessing*

1) Zero and missing values: Since a zero reading in Glucose, Blood Pressure, Skin Thickness, Insulin, or BMI is not physiologically possible, such entries were treated as missing and addressed ahead of modelling.

2) Feature scaling: Continuous variables went through z-score normalization via StandardScaler so that distance-based and gradient-based classifiers alike could work with comparable feature ranges.
3) Train-test split: An 80/20 stratified split separated the data into training and held-out test sets, keeping class proportions consistent across both.
4) Cross-validation: To get a more dependable sense of how each model generalizes, a 10-fold cross-validation scheme was run on the training set.

C. Classification Algorithms

Eight classifiers went through the same preprocessing pipeline for a fair comparison: Logistic Regression (a linear probabilistic model); Decision Tree (recursive, rule-based partitioning); Random Forest [18], a bagging ensemble of trees; Support Vector Machine [19] with an RBF kernel; K-Nearest Neighbors with k=11; Gaussian Naïve Bayes, which assumes independence between features; AdaBoost [20], a sequential boosting method; and Gradient Boosting, an additive stage-wise boosting approach. Hyperparameters drew on values commonly reported in prior diabetes-prediction work [3, 4, 8], then were fine-tuned through trial and error, and the entire pipeline ran on Scikit-learn [21].

3.4 Evaluation Metrics

Five metrics anchored the evaluation: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Precision captures the share of predicted-positive cases that were truly diabetic, while Recall (sensitivity) tells us what fraction of actual diabetic cases the model managed to catch — a number that matters clinically, since a missed diagnosis (false negative) tends to cost far more than a false alarm in a screening setting. AUC-ROC rounds this out by measuring discriminative power independent of any particular threshold choice.

IV. RESULTS AND DISCUSSION

The held-out test performance of all eight classifiers, together with 10-fold cross-validation accuracy on the training data, is laid out in Table 2.

Table 2: Comparative Performance of Machine Learning Classifiers

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	10-Fold CV Acc.
Gradient Boosting	0.7532	0.6600	0.6111	0.6346	0.8257	0.7427
Decision Tree	0.7468	0.7143	0.4630	0.5618	0.7905	0.7118
AdaBoost	0.7468	0.6667	0.5556	0.6061	0.8166	0.7476
Random Forest	0.7468	0.6744	0.5370	0.5979	0.8122	0.7688
SVM (RBF Kernel)	0.7403	0.6522	0.5556	0.6000	0.7964	0.7721
K-Nearest Neighbors	0.7338	0.6327	0.5741	0.6019	0.7867	0.7509
Logistic Regression	0.7078	0.6000	0.5000	0.5455	0.8130	0.7802
Naïve Bayes	0.7013	0.5667	0.6296	0.5965	0.7646	0.7573

Table 2 puts Gradient Boosting at the top on three fronts at once — test accuracy (75.32%), AUC-ROC (0.8257), and F1-score (0.6346) — which speaks well of stage-wise boosting for this dataset. AdaBoost and Random Forest trail only slightly behind at 74.68% accuracy, another sign that tree-based ensembles suit structured clinical data. Decision Tree, meanwhile, wins on precision (71.43%) but pays for it with recall of just 46.30% — a conservative model that avoids false alarms while letting a fair number of true diabetic cases slip through. Logistic Regression tells a different story: its test accuracy (70.78%) looks unimpressive next to the others, yet its 10-fold cross-validation accuracy (78.02%) is the best of the group, hinting that it generalizes better than the single test split suggests.

Naïve Bayes sits at the bottom on test accuracy (70.13%) but at the top on recall (62.96%), catching more true diabetic cases than any other model here, albeit with a higher false-positive count. This is a meaningful trade-off in a clinical sense: for screening purposes, a recall-favouring model like Naïve Bayes can still be the better choice even with weaker overall accuracy, given that missing an at-risk patient usually carries a steeper cost than a false alarm.

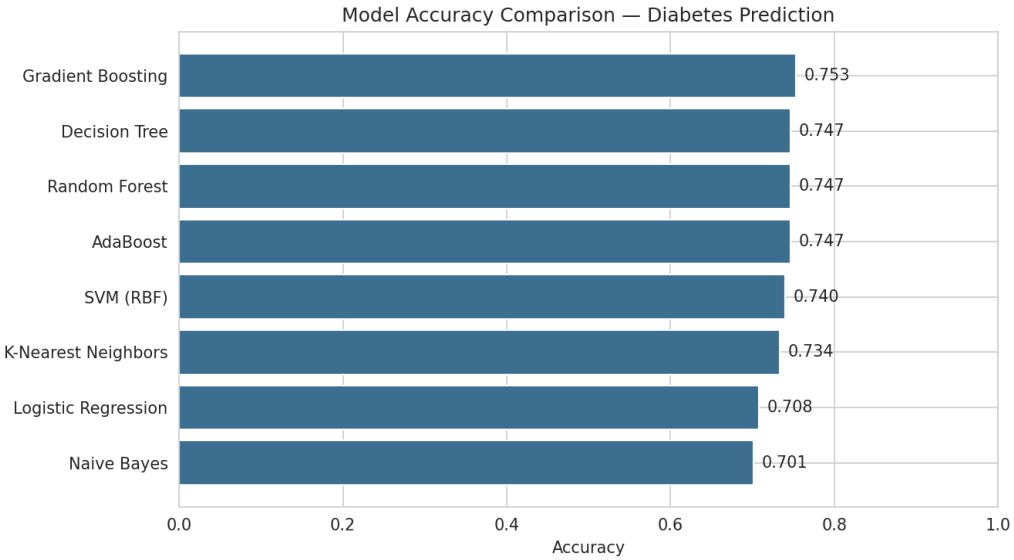


Figure 1: Test Accuracy Comparison Across Classifiers

The Random Forest model's feature-importance output, plotted in Figure 2, points to which clinical attributes matter most for predicting diabetes in this dataset.

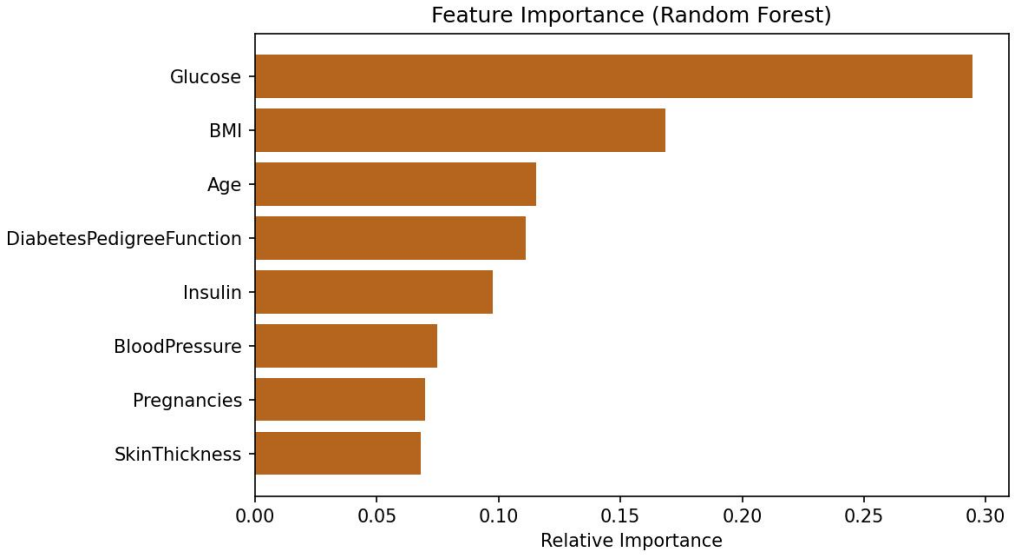


Figure 2: Feature Importance Derived from Random Forest

Table 3: Feature Importance Ranking (Random Forest)

Feature	Relative Importance
Glucose	0.2946
BMI	0.1686
Age	0.1153
Diabetes Pedigree Function	0.1111
Insulin	0.0976
Blood Pressure	0.0749
Pregnancies	0.0699
Skin Thickness	0.0680

Topping the list was glucose concentration (importance = 0.2946), with BMI (0.1686) and Age (0.1153) not far behind. None of this comes as a surprise clinically: glucose readings, whether from glycosylated hemoglobin or fasting/postprandial tests, remain the go-to diagnostic marker for diabetes, and both BMI and age are long-recognized Type 2 diabetes risk factors — a pattern also seen in Chao and Liu's work [9]. Seeing the model's rankings line up so closely with known pathophysiology lends extra weight to its clinical credibility.

No one model wins on every count here: Gradient Boosting takes the single train-test split, Naïve Bayes wins on recall and thus screening usefulness, and Logistic Regression comes out ahead across cross-validation folds. This spread is not out of line with earlier work on the same dataset, where classical models like Logistic Regression, SVM, and Decision Tree have reported accuracies from 83.5% up to 90.3% [3, 4, 8], and neural networks have touched 88.6% depending on preprocessing choices [4]. Deep architectures [13,14,15] clearly can push performance higher under the right conditions, but well-tuned ensemble and probabilistic classifiers hold their own while staying far easier to interpret and deploy. That balance — between accuracy, recall, and interpretability — matters a great deal for real clinical decision-support systems, where practitioners need to be able to trust and explain what the model is telling them.

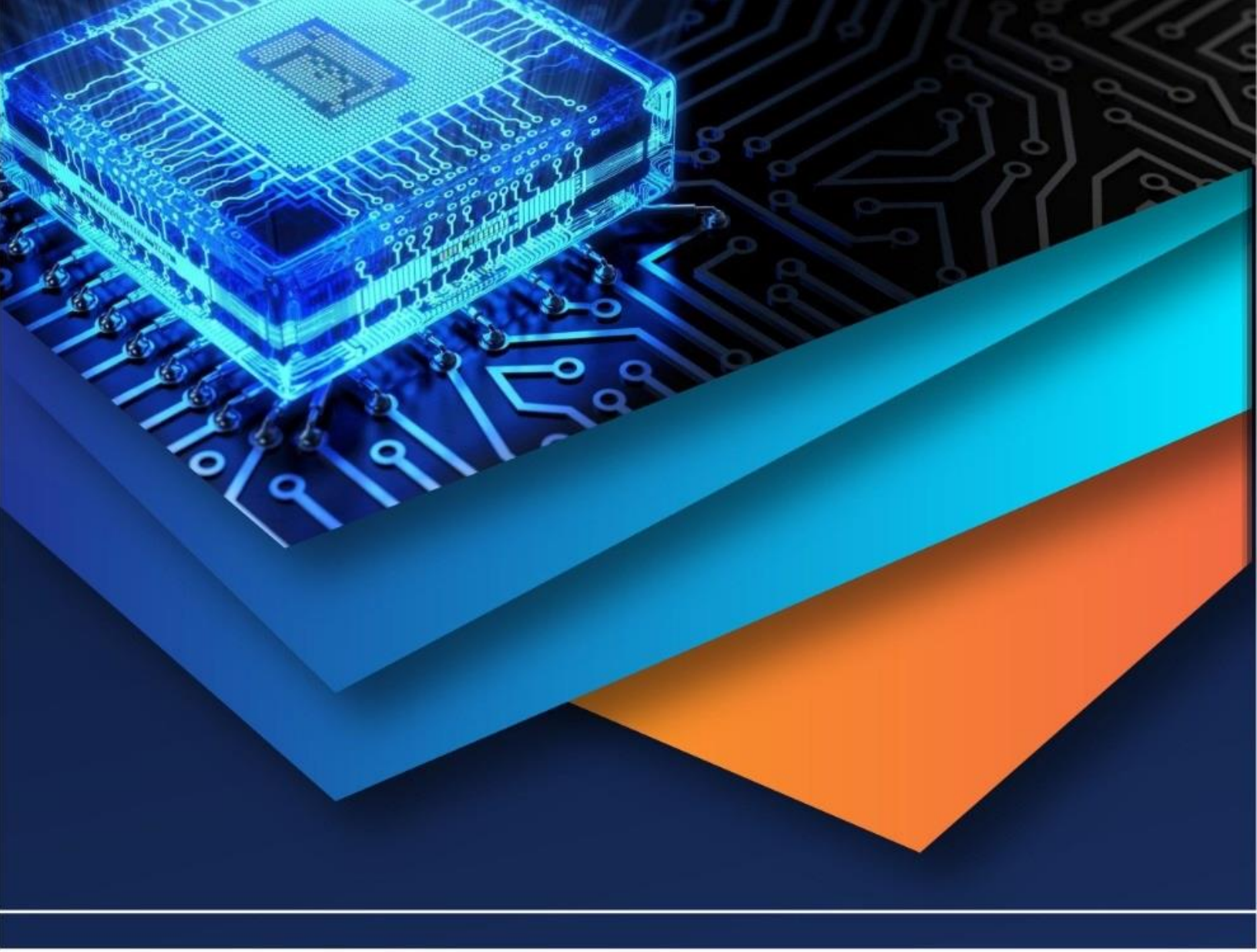
V. CONCLUSION AND FUTURE SCOPE

Across this comparison of eight machine learning algorithms on the Pima Indians Diabetes Dataset, run through one consistent evaluation pipeline, Gradient Boosting emerged ahead of the pack on both accuracy and AUC-ROC, with AdaBoost and Random Forest not far off; Glucose, BMI, and Age stood out as the most clinically relevant predictors per the Random Forest feature-importance analysis. Together, these findings make a reasonable case for boosting-based ensembles as a foundation for interpretable, accurate diabetes screening — the kind that could realistically fit into low-cost clinical decision-support systems or mobile health apps.

A few directions could take this further: larger and more demographically varied datasets, SMOTE-based class balancing [10] to handle imbalance, hybrid CNN-LSTM architectures [13,14,15] for capturing more complex, non-linear relationships among clinical variables, and explainability tools like SHAP or LIME for instance-level interpretability along the lines of Chao and Liu's work [9]. Beyond that, building this framework into an actual mobile or web-based screening tool would be a natural next step toward real healthcare impact.

REFERENCES

[1] World Health Organization. (2024). Diabetes Fact Sheet. WHO, Geneva.
[2] Pima Indians Diabetes Dataset. Kaggle / UCI Machine Learning Repository, National Institute of Diabetes and Digestive and Kidney Diseases.
[3] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585.
[4] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432–439.
[5] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292–299.
[6] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10, 11981.
[7] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.
[8] Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. (2023). Diagnostics, 13(14), 2383.
[9] Chao, Z., & Liu, J. (2026). Machine Learning-Based Diabetes Prediction Model Construction and Analysis: A Case Study of the Kaggle Diabetes Dataset. In Optimization and Data Science in Industrial Engineering (ODSIE 2025), Communications in Computer and Information Science, vol. 2854. Springer, Cham.
[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
[11] Akula, R., Nguyen, N., & Garibay, I. (2019). Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes. IEEE SoutheastCon 2019, 1–9.
[12] Patro, S., Sharma, V., & Vishwakarma, S. (2023). A comparative study of machine learning models for diabetes prediction. BMC Bioinformatics, 24, 372.
[13] Aslan, M. F., & Sabanci, K. (2023). A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data. Diagnostics, 13(4), 796.
[14] Fan, Y. (2025). Diabetes diagnosis using a hybrid CNN-LSTM-MLP ensemble. Scientific Reports, 15, 26765.
[15] Enhanced diabetes prediction using pre-trained CNNs, LSTM, and conditional GAN on transformed numerical data. (2026). Scientific Reports.
[16] Farnoosh, R., et al. (2025). DiabetesXpertNet: An innovative attention-based CNN for accurate type 2 diabetes prediction. PLOS ONE.
[17] Noh, M. J., & Kim, Y. S. (2025). Diabetes Prediction Through Linkage of Causal Discovery and Inference Model with Machine Learning Models. Biomedicines, 13(1), 124.
[18] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
[19] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
[20] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.
[21] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY