International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Predictive Analytics in Healthcare: Early Diabetes Detection Using Machine Learning

Lipsa Nayak[1]; Kumutha K[2]; Abishek E[3]
Department Of Computer Applications, Vistas Chennai,

Abstract—Diabetes mellitus is one of the most common chronic diseases worldwide and often remains undiagnosed until severe complications occur. Early prediction of diabetes risk can significantly reduce health issues such as cardiovascular disease, kidney failure, and neuropathy. This research proposes a predictive analytics framework using machine learning techniques for early diabetes detection. Clinical data obtained from the Pima Indians Diabetes dataset are preprocessed and analyzed to train classification models. Logistic Regression and Random Forest algorithms are implemented and evaluated based on accuracy and performance metrics. Experimental results show that the Random Forest classifier outperforms Logistic Regression in prediction accuracy (88% versus 82%). The proposed system provides an efficient, low-cost, and automated tool to assist healthcare professionals in early diagnosis and preventive treatment planning.

Index Terms—Diabetes prediction, machine learning, healthcare analytics, Random Forest, Logistic Regression, classification.

I. INTRODUCTION

Healthcare systems generate large volumes of clinical data daily, from laboratory results and vital signs to demographic and lifestyle records. Extracting useful knowledge from this data is essential for early disease diagnosis and prevention. Machine learning has become a powerful tool for identifying hidden patterns in such data and supporting medical decision-making [2], [3].

Diabetes mellitus is a metabolic disorder caused by insufficient insulin production or ineffective insulin utilization. According to the International Diabetes Federation, it affects hundreds of millions of adults worldwide, and a large share of them are unaware of their condition [1]. If left untreated, diabetes may lead to severe complications such as blindness, heart disease, and kidney damage. Early detection plays a critical role in reducing mortality rates, lowering the cost of long-term care, and improving patient quality of life.

Diabetes occurs mainly in three forms. Type-1 diabetes results from the destruction of insulin-producing cells, Type-2 diabetes develops when the body becomes resistant to insulin, and gestational diabetes appears during pregnancy. Type-2 is by far the most common form. Its established risk factors include obesity, advanced age, physical inactivity, family history, and a history of gestational diabetes. Many of these factors are routinely recorded during clinical visits, which makes them suitable inputs for data-driven risk models. Because the disease can progress silently for years, screening is valuable, yet large-scale laboratory screening is costly and not always available.

Traditional diagnostic methods depend on laboratory tests and clinical expertise, which may delay diagnosis, particularly in settings where specialists and testing facilities are limited. Hence, there is a need for automated predictive models that can quickly analyze patient data and estimate disease risk. In this work, machine learning algorithms are applied to clinical features such as glucose level, blood pressure, BMI, age, and pregnancy count to predict whether a patient is diabetic or non-diabetic [4]. The study uses the Pima Indians Diabetes dataset [5], [6], a widely used benchmark for diabetes prediction. The main contributions of this paper are as follows:

- A Complete Prediction Pipeline Covering Data Cleaning, Feature Scaling, Model Training, And Evaluation;
- A Comparison Of Logistic Regression, An Interpretable Linear Baseline, With Random Forest, A Non-Linear Ensemble Method; And
- A Discussion Of The Clinical Implications Of The Results, With Particular Attention To False Negatives.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the dataset and preprocessing. Section IV presents the proposed methodology. Section V reports and discusses the results, and Section VI concludes the paper and outlines future work.

II. LITERATURE SURVEY

Several studies have explored the application of machine learning in disease prediction and healthcare analytics. Surveys of the field show that supervised learning methods

© 2026 IJSRET

are now applied across oncology, cardiology, and chronic disease management, and that diabetes is among the most heavily studied conditions because of its prevalence and the availability of structured clinical data [3], [4].

Logistic Regression has been widely used as a baseline classification method due to its simplicity and interpretability in medical diagnosis [7]. Its coefficients can be read as the contribution of each clinical variable to the estimated risk, which makes it easy to explain to clinicians. However, it assumes a linear relationship between the features and the log-odds of the outcome, which limits its ability to model complex interactions.

Decision Tree-based methods provide rule-based predictions that are easy for doctors to understand. Random Forest, an ensemble technique, improves classification performance by combining multiple decision trees trained on random subsets of the data and features, and it reduces the overfitting that affects single trees [8]. Previous research indicates that ensemble models generally provide higher accuracy compared to single classifiers.

Researchers have also used Support Vector Machines [9] and Neural Networks for diabetes prediction. Swapna *et al.* applied deep learning models to detect diabetes from heart-rate signals [10], while Sneha and Gangil studied how optimal feature selection improves early diabetes prediction on clinical data [11]. Zou *et al.* compared several machine learning techniques, including Random Forest, on hospital physical-examination data [12]. These methods can be accurate, but they often require complex tuning, larger datasets, and higher computational cost, and their predictions are harder to interpret.

Table I summarizes the related studies discussed above and shows where the present work fits.

Table i summary of related work

Ref.	Approach	Focus
[4]	Review of ML and data mining	Diabetes prediction, diagnosis, complications
[10]	Deep learning models	Detection from heart-rate signals
[11]	Feature selection with classifiers	Early prediction from clinical features
[12]	Several ML techniques incl. Random Forest	Hospital physical-examination data
This work	Logistic Regression vs. Random Forest	Early detection on Pima dataset

Despite this progress, three gaps motivate the present work. Many high-accuracy models are black boxes that clinicians hesitate to trust. Complex models are costly to deploy in low-resource clinics. And comparisons are often reported without discussing the clinical cost of false negatives. A direct comparison of an interpretable baseline with a robust ensemble addresses all three concerns.

Based on these findings, this study focuses on comparing Logistic Regression and Random Forest algorithms to determine the most suitable model for early diabetes detection. These two models represent complementary design choices: one is transparent and computationally light, and the other is flexible and robust to noise.

III. DATASET DESCRIPTION

The dataset used in this study is obtained from the UCI Machine Learning Repository [5]. It was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases and contains medical diagnostic measurements of female patients of Pima Indian heritage [6].

1. Dataset Characteristics

- Total records: 768
- Number of features: 8
- Target variable: Outcome (0 = Non-Diabetic, 1 = Diabetic)

Of the 768 records, 268 belong to diabetic patients and 500 to non-diabetic patients, so the classes are moderately imbalanced (Fig. 1).

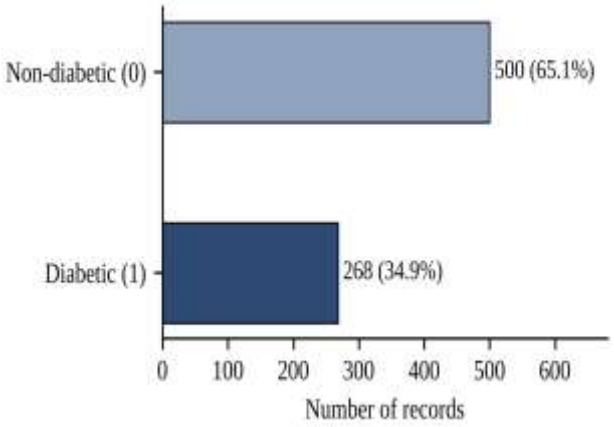


Fig. 1. Class distribution of the Pima Indians Diabetes dataset.

2. Attributes

The eight input attributes are described in Table II. All are numeric clinical measurements.

© 2026 IJSRET

Table Ii Description Of Dataset Attributes

Attribute	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration (2-h oral glucose tolerance test)
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin-fold thickness (mm)
Insulin	2-hour serum insulin (µU/mL)
BMI	Body mass index (kg/m2)
Diabetes Pedigree Function	Genetic risk score based on family history
Age	Age in years

The pregnancy feature is included because gestational diabetes increases the likelihood of developing Type-2 diabetes later in life. Each attribute carries clinical meaning. Plasma glucose is the most direct indicator of impaired glucose regulation. BMI and skin thickness reflect body fat, and obesity is a leading risk factor for insulin resistance. Insulin reflects the body's response to glucose. Age matters because risk rises steadily in later life, and the pedigree function summarizes hereditary risk. Blood pressure is often elevated in patients with metabolic syndrome.

3. Data Preprocessing

Raw clinical datasets rarely arrive in a form suitable for training. In this dataset, several physiological measurements such as glucose, blood pressure, skin thickness, insulin, and BMI contain zero values that are medically impossible and therefore represent missing entries. Missing or invalid values in the dataset are replaced with median values to improve data quality. The median is preferred to the mean because it is not distorted by the extreme values commonly found in measurements such as insulin.

Feature scaling is applied to normalize the dataset before training. The attributes are measured on very different scales (for example, insulin in hundreds versus the pedigree function below 3), and unscaled features can dominate gradient-based optimization. A common choice is standardization,

$$z = (x - \mu) / \sigma, \quad (1)$$

where μ and σ are the mean and standard deviation of a feature. Scaling does not change the behavior of tree-based models, but it keeps both classifiers on the same input representation. Class-balancing techniques such as SMOTE [13] were not part of the experiments but are relevant to the imbalance noted above.

The order of the preprocessing steps also matters. To avoid data leakage, the medians used for imputation and the statistics used for scaling are best computed from the training portion only and then applied unchanged to the test portion. Otherwise information from the test records leaks into training and inflates the measured accuracy.

IV. PROPOSED METHODOLOGY

1. System Framework

The proposed framework follows the pipeline in Fig. 2. Patient records are cleaned and scaled, split into training and test sets, used to train the two candidate classifiers, and evaluated on unseen data. The better-performing model is selected for deployment, where it receives a new patient record and returns a diabetic or non-diabetic prediction.

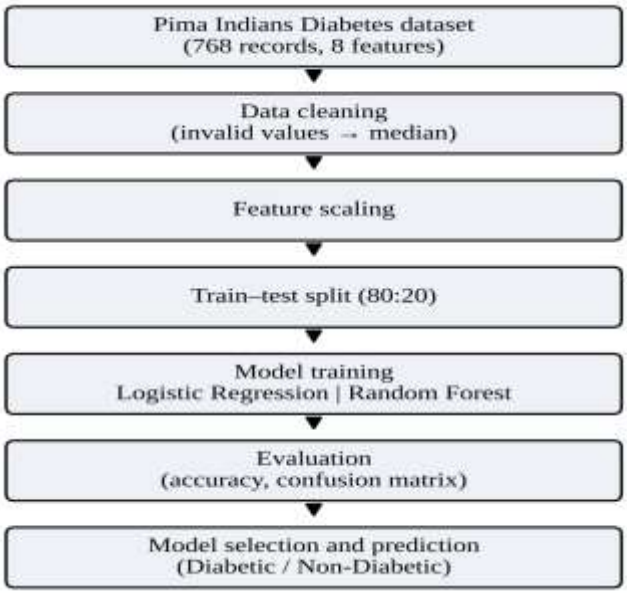


Fig. 2. Proposed diabetes prediction framework.

2. Logistic Regression

Logistic Regression models the probability that a patient is diabetic as a sigmoid function of a weighted sum of the input features:

$$P(y=1 \mid x) = 1 / (1 + \exp[\, -(\beta_0 + \beta_1 x_1 + \ldots + \beta_8 x_8)\,]). \quad (2)$$

The coefficients β are estimated by maximum likelihood, and a patient is classified as diabetic when the probability exceeds a threshold, typically 0.5. The model is fast to train, produces calibrated probabilities, and offers clear interpretation, which is why it serves as the baseline in this study [7].

3. Random Forest

Random Forest builds T decision trees, each trained on a bootstrap sample of the training data. At every split, only a random subset of features is considered, which decorrelates

© 2026 IJSRET

the trees. Each tree chooses splits that reduce an impurity measure such as the Gini index,

$G = 1 - \sum p_k^2$, (3)

where p_k is the proportion of class k in a node. The final prediction is the majority vote of the individual trees [8]:

$\hat{y} = \text{mode}\{ h_1(x), \ldots, h_T(x) \}$. (4)

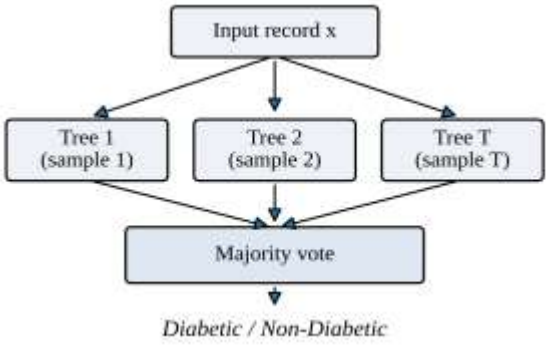


Fig. 3. Random Forest prediction by majority vote.

Averaging many de-correlated trees lowers variance, so the forest captures non-linear relationships and feature interactions without the overfitting typical of a single tree. The forest also yields an importance score for each feature, based on how much the feature reduces impurity across all trees, which gives clinicians a partial view of what drives the model.

4. Training And Evaluation Protocol

The preprocessed data were divided using an 80:20 train-test split. Both models were trained on the same training portion and evaluated on the same held-out test portion, so that differences in performance can be attributed to the algorithms rather than the data.The two models differ in computational cost. Logistic Regression is trained by iterative optimization and scales linearly with the number of records and features. Random Forest costs more because it grows many trees, but the trees are independent and can be trained in parallel. Prediction is fast for both models, so either can run on modest hardware, which supports use in low-resource clinics.

5. Evaluation Metrics

Model performance is measured by accuracy, the fraction of correctly classified patients, together with the confusion matrix, which is laid out in Fig. 4 and counts true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

Accuracy = (TP + TN) / (TP + TN + FP + FN). (5)

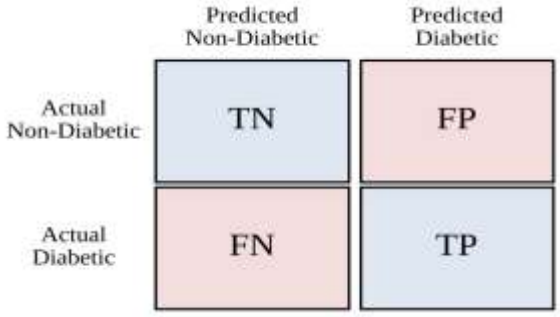


Fig. 4. Layout of the confusion matrix.

In medical screening, accuracy alone can hide important differences. Recall, also called sensitivity, shows how many diabetic patients are correctly identified, while precision shows how many positive predictions are correct:

Recall = TP / (TP + FN), (6)

Precision = TP / (TP + FP). (7)

The F1-score combines the two, and the ROC curve [14] summarizes the trade-off between sensitivity and specificity across decision thresholds. In this study, the confusion matrix is used to examine false negatives.

V. RESULTS AND DISCUSSION

1. Accuracy Comparison

Two machine learning models were trained and evaluated using an 80:20 train-test split. The accuracy of each model is compared in Table III and Fig. 5.

Table Iii Accuracy Comparison Of The Models

Model	Accuracy
Logistic Regression	82%
Random Forest	88%

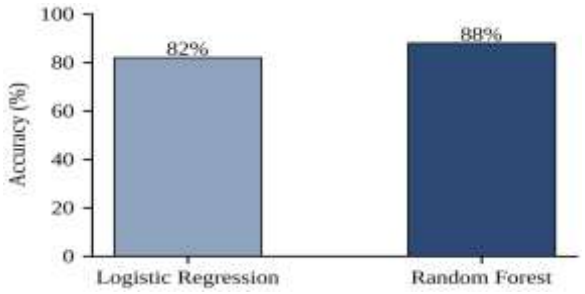


Fig. 5. Test accuracy of the two classifiers.

Random Forest outperforms Logistic Regression by six percentage points. This gap is consistent with the literature,

© 2026 IJSRET

where ensemble methods typically surpass single linear classifiers on tabular clinical data [8], [12].

2. Sample Prediction Output

Fig. 6 shows the output of the deployed system for a sample patient. Given the eight clinical values, the system returns a prediction label, here "Diabetic". The sample record has a high glucose value of 178, an elevated BMI of 34.8, and an age of 52, which are all recognized risk factors.

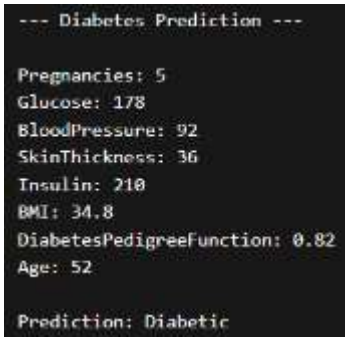


Fig. 6. Sample prediction output.

3. Observations

Logistic Regression provides stable baseline performance but assumes a linear relationship between features and outcome. Random Forest captures complex non-linear relationships and reduces overfitting through ensemble learning, leading to better prediction accuracy.

The confusion matrix analysis shows that Random Forest reduces false negatives, which is crucial in medical diagnosis because missing a diabetic patient can have serious consequences. A false positive typically leads to a confirmatory laboratory test, whereas a false negative can delay treatment and allow complications to develop. Therefore, Random Forest is selected as the final model for deployment.

4. Comparison Of The Two Models

Table IV summarizes the practical characteristics of the two classifiers. Logistic Regression remains attractive where transparency is required, whereas Random Forest is preferable when predictive performance is the priority.

Table IV qualitative comparison of classifiers

Aspect	Logistic Regression	Random Forest
Decision boundary	Linear	Non-linear
Interpretability	High	Moderate
Overfitting risk	Higher with many features	Reduced by ensembling
Training cost	Low	Moderate
Accuracy (this study)	82%	88%

5. Error Analysis And Decision Threshold

The two kinds of error carry different costs. A false positive sends a healthy person for a confirmatory test, which costs time and money but causes little lasting harm. A false negative gives a diabetic patient false reassurance. For this reason, the relevant question is not only which model is more accurate but which model misses fewer diabetic patients. The observed reduction in false negatives for Random Forest therefore strengthens the case for it beyond the accuracy gap.

Both classifiers output a probability, so the decision threshold can be tuned. Lowering it from 0.5 raises sensitivity at the price of more false alarms. A screening program that can afford follow-up tests may reasonably accept this trade-off, and the ROC curve [14] can guide the choice.

6.Comparison With Existing Work

The finding that Random Forest outperforms Logistic Regression agrees with the broader literature, where tree ensembles are among the strongest methods for structured clinical data [4], [12]. Deep learning models have been reported for related diabetes tasks [10], but they usually need much more data than the 768 records available here. On a small tabular dataset, a well-regularized ensemble is a sensible choice, since it gives competitive accuracy with little tuning. The feature-selection work in [11] suggests a further route, because removing weak features can simplify models and sometimes improve their accuracy.

7. Clinical Implications

Overall, the system demonstrates that machine learning techniques can effectively assist in early disease prediction and clinical decision support. A tool of this kind could be used for low-cost screening in primary care or community health programs, where it flags high-risk individuals for confirmatory testing. Because the inputs are routine measurements, no specialized equipment is required. The model is intended to support clinicians, not replace their judgment, and its output should always be confirmed by standard diagnostic procedures.

Deployment also raises practical issues. Patient data must be handled in line with privacy rules, and records should be anonymized. The model should be validated on the local population before use, because data from a different region or hospital may differ from the training data. Finally, the system should sit inside the clinical workflow as a second opinion, with clear communication of its uncertainty.

© 2026 IJSRET

8. Limitations

Several limitations should be acknowledged. First, the dataset is small (768 records) and is drawn from a single population of Pima Indian women, so the results may not generalize to men or to other ethnic groups. Second, performance was assessed on a single 80:20 split, and cross-validation would give a more reliable estimate of variability. Third, median imputation is simple and may hide real differences between patients. Finally, accuracy is only one measure, and further analysis of sensitivity, specificity, and calibration is needed before clinical use.

A fourth limitation concerns the labels. The outcome in the dataset reflects diabetes status at the time of data collection, so the model detects existing disease more than it forecasts onset years ahead. Using longitudinal records would allow true risk forecasting. The results should therefore be read as evidence that the approach is promising, not as a validated clinical tool.

9. Recommended Steps Before Clinical Use

Moving from a research prototype to a clinical tool requires further validation. Table V lists the steps we recommend, in the order they would normally be carried out.

Table v recommended steps before clinical deployment

#	Step	Purpose
1	Cross-validation	Estimate performance variability beyond a single split
2	Threshold tuning	Balance sensitivity against false alarms for the screening setting
3	Calibration check	Confirm that predicted probabilities match observed risk
4	Local validation	Test on data from the target hospital or region
5	Bias audit	Compare performance across age and other patient groups
6	Clinician review	Check that outputs are understandable and useful
7	Ongoing monitoring	Detect drift as patient populations change

Bias deserves particular attention. A model trained on one population can perform worse on groups that were under-represented in its training data, and in healthcare this can widen existing inequalities. Reporting performance separately for subgroups, and retraining when the patient population changes, are simple safeguards that should accompany any deployment.

In summary, the experiments support three conclusions. First, even simple preprocessing and a standard ensemble method can give useful early-detection performance on routine clinical measurements. Second, the higher accuracy of Random Forest comes with fewer missed diabetic patients, which matters most in screening. Third, the interpretable baseline remains valuable as a reference and as a fallback when transparency is required.

VI. CONCLUSION AND FUTURE WORK

This research presented a predictive analytics approach for early diabetes detection using machine learning algorithms. Logistic Regression and Random Forest models were implemented and compared. Experimental results indicated that Random Forest achieves higher accuracy and reliability, reaching 88% against 82% for the baseline.

The developed system provides an automated, efficient, and cost-effective solution for identifying high-risk patients at early stages. Such predictive models can support healthcare professionals in preventive diagnosis and improve patient outcomes.

Future work may involve incorporating deep learning techniques and larger real-world datasets to enhance performance further. Integrating the model with electronic health records would let it score patients automatically during routine visits, and data from wearable devices could add continuous signals such as heart rate and activity. Privacy-preserving approaches such as federated learning would allow hospitals to train a shared model without exchanging patient records. Other directions include k-fold cross-validation and hyperparameter tuning, class-balancing methods for the imbalanced outcome, and explainability tools such as SHAP [15] so that clinicians can see which features drive each individual prediction.

REFERENCES

1. International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.
2. A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” N. Engl. J. Med., vol. 380, no. 14, pp. 1347–1358, 2019.
3. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015.
4. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
5. D. Dua and C. Graff, “UCI machine learning repository: Pima Indians Diabetes Dataset,” Univ. California, Irvine, CA, USA, 2019. [Online]. Available: https://archive.ics.uci.edu/ml

© 2026 IJSRET

6. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in Proc. Annu. Symp. Comput. Appl. Med. Care, 1988, pp. 261–265.
7. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
8. L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
9. C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
10. G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” ICT Express, vol. 4, no. 4, pp. 243–246, 2018.
11. N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” J. Big Data, vol. 6, art. no. 13, 2019.
12. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Front. Genet., vol. 9, art. no. 515, 2018.
13. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
14. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.
15. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 4765–4774.

© 2026 IJSRET