Comparative Evaluation of Logistic Regression and Random Forest for Diabetes Risk Prediction using the DiaBD Clinical Dataset

Okoro Favour Ngozi
Data Analytics
Department of Computer Science,
Ebonyi State University, Ebonyi State, Nigeria

Ituma Chinagorom
Artificial intelligence, Automation and Robotics
Department of Computer Science,
Ebonyi State University, Ebonyi State, Nigeria

Chinonso Job
Computer Science
School of Built Environment, Engineering, and Computing
LEEDSBECKETT University

ABSTRACT
Diabetes mellitus remains one of the fastest-growing chronic diseases worldwide, and early risk identification is essential for timely clinical intervention. This paper presents a comparative evaluation of two supervised machine learning algorithms, Logistic Regression (LR) and Random Forest (RF), for diabetes risk classification using the DiaBD dataset, a recently published clinical dataset of 5,288 patient records from Bangladesh. Chi-square and analysis-of-variance (ANOVA) feature selection identified six clinically relevant predictors: hypertension status, glucose level, systolic and diastolic blood pressure, age, and weight. The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training partition only, to correct for class imbalance without leaking information into the test set. Both models were trained on an 80/20 train-test split and evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Random Forest achieved a higher overall accuracy (0.88) and a marginally higher AUC (0.83) than Logistic Regression (0.81 accuracy, 0.82 AUC); Logistic Regression, however, achieved substantially higher recall for the diabetic class (0.66 versus 0.44). The two models' F1-scores were nearly identical (0.32 for both), which shows that a single blended metric conceals the clinically important difference in how each model distributes its errors. Precision remained modest for both models (0.21-0.25), indicating that a six-variable feature set is not sufficient for high-precision clinical deployment on its own. We argue that model selection for diabetes screening should be guided by the clinical cost of missed diagnoses rather than accuracy alone, and we recommend threshold calibration, cost-sensitive learning, and external validation on additional populations as directions for future work.

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords
Diabetes prediction, Logistic Regression, Random Forest, machine learning, SMOTE, class imbalance, healthcare decision support, DiaBD dataset

1. INTRODUCTION
Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycaemia, resulting from insufficient insulin production or the body's inability to use insulin effectively [1]. The global burden of diabetes continues to rise: the World Health Organization reports that 14% of adults aged 18 years and above were living with diabetes in 2022, up from 7% in 1990 [1], and it remains among the leading causes of death and disability-adjusted life years worldwide [2], [3]. Early identification of at-risk individuals is central to reducing the long-term complications of the disease, including cardiovascular disease, nephropathy, and retinopathy.

Traditional diagnostic pathways rely on laboratory testing and clinician judgement, which, while accurate, are resource- and time-intensive and are not well suited to large-scale, low-cost screening, particularly in low- and middle-income countries. Machine learning (ML) offers a complementary, data-driven approach: by learning patterns from historical patient records, ML models can flag high-risk individuals for further clinical evaluation before symptoms become severe.

A large body of prior work has applied ML to diabetes prediction, most commonly using the Pima Indians Diabetes (PID) dataset, and has reported accuracies ranging from roughly 75% to above 97% depending on the algorithm, feature engineering strategy, and dataset used [4]-[10]. Three recurring gaps motivate the present study. First, many studies report only a single performance metric, typically accuracy, which can be misleading on imbalanced medical datasets where the minority (positive) class is the class of clinical interest. Second, relatively few studies report results on datasets from outside a small number of well-studied Western or South Asian cohorts, which limits insight into how well these models generalize. Third, trade-offs between competing objectives, such as overall accuracy versus sensitivity to the disease-positive class, are often understated in favour of headline accuracy figures.

This paper addresses these gaps by presenting a transparent, multi-metric comparison of Logistic Regression and Random Forest on the DiaBD dataset [11], a clinical dataset of 5,288 patient records recently collected in Bangladesh and made publicly available via Mendeley Data. We report accuracy, precision, recall, F1-score, and AUC-ROC for both models, and we explicitly discuss the accuracy/recall trade-off that emerged, rather than declaring a single "winner" based on accuracy alone. The specific contributions of this paper are:

i. A reproducible feature selection pipeline using chi-square and ANOVA tests to identify the most predictive clinical variables for diabetes risk in the DiaBD dataset;

ii. A leakage-safe application of SMOTE, applied strictly to the training partition, to address class imbalance;

iii. A multi-metric comparison of Logistic Regression and Random Forest, including a harmonic-mean (F1) analysis that exposes a trade-off the headline accuracy and AUC figures do not show on their own; and

iv. A candid discussion of the precision limitations of both models, the single-dataset scope of this study, and concrete

recommendations for improving clinical deployability and external validity.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, preprocessing pipeline, and the two learning algorithms. Section 4 presents and discusses the results. Section 5 concludes the paper and outlines future work.

2. RELATED WORK

Machine learning approaches to diabetes prediction have been extensively studied, predominantly using the Pima Indians Diabetes dataset. Khanam and Foo [4] compared several classical ML algorithms and neural networks on the PID dataset, finding that Logistic Regression and Support Vector Machines performed competitively, with a two-hidden-layer neural network reaching 88.6% accuracy. Hasan et al. [5] combined outlier rejection, feature selection, and an AUC-weighted ensemble classifier to achieve an AUC of 0.95 on the same dataset. Olisah et al. [6] applied Spearman correlation-based feature selection with a two-gradient-descent neural network, reporting 97.25% accuracy.

Ensemble and hybrid approaches have also shown strong results. Zhou et al. [7] proposed a Boruta feature selection and ensemble learning pipeline (DPMBFSEL) that achieved 98% accuracy on the PID dataset, while Doğru et al. [8] introduced a super-learner model combining Logistic Regression, decision trees, Random Forest, and gradient boosting as base learners with an SVM meta-learner, improving early-diagnosis accuracy to 92% across three datasets. Alghamdi [9] applied XGBoost to high-dimensional feature sets and reported 89% accuracy, while Ahmed et al. [10] combined SVM and artificial neural networks in a fused model (FMDP), reaching 94.87% accuracy.

Deep learning has also been explored. García-Ordás et al. [13] combined variational and sparse autoencoders with convolutional neural networks to reach 92.31% accuracy with oversampling and feature augmentation, while Bukhari et al. [14] used a back-propagation scaled-conjugate-gradient neural network to achieve 93% accuracy on the PID dataset. Zhao et al. [15] proposed an attention-oriented convolutional network validated on both the NHANES and PID datasets, achieving 89.47% and 94.12% accuracy respectively.

A smaller body of work has moved beyond the PID dataset. Deberneh and Kim [11] used a large hospital-derived Korean cohort of over 253,000 records with SMOTE-based imbalance correction, reporting an F1-score of 74% for both Random Forest and SVM classifiers, notably lower than PID-based studies, an early indicator that performance figures obtained on the small, extensively studied PID dataset may not transfer to larger, more heterogeneous clinical populations. Tasin et al. [12] combined the PID dataset with a private Bangladeshi textile-industry cohort and applied explainable AI (SHAP, LIME) alongside SMOTE and ADASYN oversampling to improve both performance and interpretability.

Collectively, this literature demonstrates that ML-based diabetes prediction is a mature and active research area, but it also reveals two consistent limitations that this paper addresses directly: a heavy concentration of results on a single, small, and demographically narrow dataset (PID, 768 records from Pima Indian women), and an emphasis on accuracy as the primary reported metric, which can obscure poor sensitivity to the disease-positive class on imbalanced data. The DiaBD dataset [16] used in this study is more than six times larger than PID and was collected from a distinct South Asian population with a broader demographic profile, providing a useful additional benchmark for the field. Like most studies in this literature, however, the present paper still evaluates its models on a single dataset; Section 4.5 returns to this limitation directly and sets out what a multi-dataset extension of this work would require.

3. MATERIALS AND METHODS

3.1 Dataset

This study uses the DiaBD dataset [16], published by Prama et al. (2025) and made publicly available through Mendeley Data. The dataset contains 5,288 patient records collected in Bangladesh, with 14 independent attributes spanning demographic (age, gender), anthropometric (height, weight, BMI), clinical (pulse rate, systolic and diastolic blood pressure, glucose level), and family/medical history variables (hypertension, cardiovascular disease, stroke). Each record carries a binary target label, Diabetic (Yes/No), indicating diagnosed diabetes status.

3.2 Data Preprocessing

Categorical variables (gender, diabetic status) were label-encoded to binary numeric values (Male = 1, Female = 0; diabetic = 1, non-diabetic = 0). The dataset was then partitioned into an 80% training set (4,230 records) and a 20% held-out test set (1,058 records) using stratified sampling to preserve class proportions, following the commonly used split ratio recommended for small-to-medium tabular datasets [17].

3.3 Feature Selection

To identify the variables most strongly associated with diabetes status, chi-square tests of independence were applied to categorical features and ANOVA F-tests to continuous features, using a significance threshold of $p \leq 0.05$. Six variables, hypertension status, glucose level, systolic blood pressure, diastolic blood pressure, age, and weight, were retained as the most predictive features and used to train both models. BMI, pulse rate, cardiovascular disease, and stroke history were found to be moderately predictive, while gender, family history of hypertension, and height were statistically insignificant ($p > 0.05$) and were excluded.

3.4 Class Imbalance Correction

Diabetic cases were substantially outnumbered by non-diabetic cases in the raw dataset. To avoid models trivially predicting the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) [18] was applied. SMOTE was fit and applied only to the training partition, after the train/test split, so that the held-out test set retained its original, real-world class distribution and no synthetic information leaked into model evaluation.

3.5 Model Development

1) Logistic Regression: Logistic Regression models the probability that a patient belongs to the diabetic class as a sigmoid function of a linear combination of the selected input features, with parameters estimated by maximizing the likelihood of the observed labels (equivalently, minimizing binary cross-entropy loss). It was implemented using scikit-learn's LogisticRegression class with max_iter = 1000 and random_state = 42, trained on the SMOTE-balanced training data. Logistic Regression was included primarily for its interpretability: its coefficients provide a direct, clinically explainable measure of each feature's contribution to diabetes risk [19].

2) Random Forest: Random Forest is an ensemble method that aggregates the predictions of many decision trees, each trained

on a bootstrap sample of the training data with a random subset of features considered at each split, with the final classification determined by majority vote [20]. It was implemented using scikit-learn's RandomForestClassifier with random_state = 42, trained on the same SMOTE-balanced training data. Random Forest was included for its capacity to capture non-linear interactions between clinical variables, for example the combined effect of elevated glucose and elevated BMI, that a linear model cannot represent directly.

3.6 Evaluation Metrics

Model performance was assessed on the untouched 1,058-record test set using accuracy (the proportion of all cases correctly classified), precision (the proportion of predicted-diabetic cases that were truly diabetic), recall/sensitivity (the proportion of true diabetic cases correctly identified), F1-score (the harmonic mean of precision and recall), and the area under the receiver operating characteristic curve (AUC-ROC), which summarizes the trade-off between sensitivity and specificity across all classification thresholds. In a clinical screening context, recall is of particular importance, since a false negative represents a missed diabetes diagnosis.

4. RESULTS AND DISCUSSION

4.1 Overall Performance

Table 1 summarizes the performance of both models on the 1,058-record held-out test set, including F1-score, which was omitted from the summary table in the earlier draft of this paper despite being one of the five metrics used throughout the study. Random Forest achieved the higher overall accuracy (0.88 vs. 0.81) and a marginally higher AUC (0.83 vs. 0.82), indicating slightly stronger overall discriminative ability across all decision thresholds. At the default 0.5 classification threshold, Logistic Regression achieved substantially higher recall for the diabetic class (0.66 vs. 0.44), while Random Forest achieved marginally higher precision (0.25 vs. 0.21).

Table 1. Model performance on the DiaBD test set (n = 1,058). AUC: LR = 0.82, RF = 0.83.

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.81	0.21	0.66	0.32
Random Forest	0.88	0.25	0.44	0.32

4.2 Class-Level Behaviour

Figures 1 and 2 show the normalized confusion matrices for Logistic Regression and Random Forest. Random Forest correctly identified 91% of non-diabetic cases but only 44% of true diabetic cases, meaning that 56% of actual diabetic patients in the test set would have been missed had Random Forest alone been used for screening. Logistic Regression correctly identified a smaller proportion of non-diabetic cases (83%) but a considerably larger proportion of diabetic cases (66%), at the cost of a higher false-positive rate (17% vs. 9%).

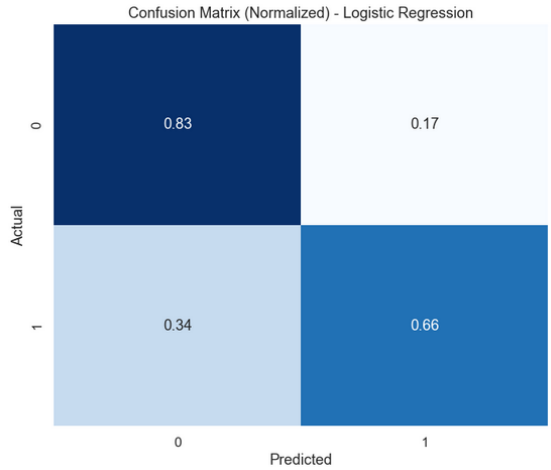


Figure 1. Normalized confusion matrix, Logistic Regression.

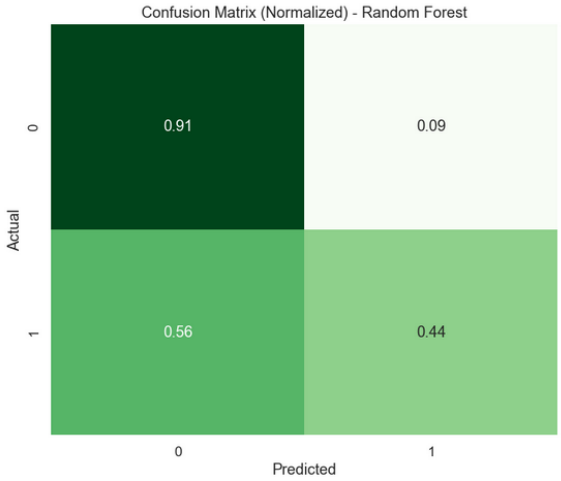


Figure 2. Normalized confusion matrix, Random Forest.

4.3 ROC Analysis

Figure 3 plots the ROC curves for both models. The two curves are close together across most of the false-positive-rate range and cross near FPR = 0.2-0.3, which is consistent with the near-identical AUC values (0.82 for Logistic Regression, 0.83 for Random Forest) and shows that neither model dominates the other across all classification thresholds. Random Forest's curve sits marginally higher through the low-to-mid false-positive-rate region, which is the region most relevant to a screening tool that must keep false alarms manageable; this gives Random Forest a modest edge in ranking ability even though, at the single 0.5 threshold evaluated in Table 1, Logistic Regression's recall is the higher of the two. This distinction, between threshold-independent ranking ability (AUC) and the sensitivity obtained at one specific operating point, is lost if only a single threshold's metrics are reported, and it is one reason threshold calibration is listed as future work in Section 5.

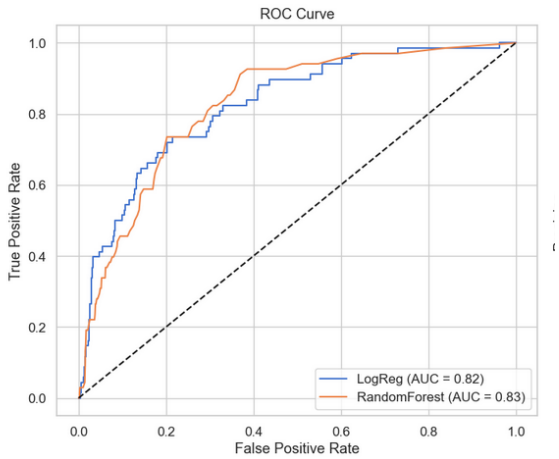


Figure 3. ROC curves for Logistic Regression and Random Forest on the DiaBD test set.

4.4 Interpretation and the Accuracy-Recall Trade-off

This asymmetry is a direct consequence of how the two models responded to the class-imbalanced but SMOTE-corrected training data, and it has a concrete clinical interpretation. Accuracy on an imbalanced test set, in which non-diabetic cases substantially outnumber diabetic cases, is dominated by performance on the majority class; a model can achieve a high accuracy score while still failing to detect a large share of true positive cases. Random Forest's higher accuracy is therefore partly a reflection of its stronger performance on the majority (non-diabetic) class rather than evidence of superior detection of diabetes itself.

The F1-scores in Table 1 add a further layer to this picture that accuracy and AUC alone do not show: both models land at an almost identical F1 of 0.32, despite having very different precision-recall balances. A single blended F1 figure therefore cannot distinguish a model that misses more diabetic cases (Random Forest) from one that raises more false alarms (Logistic Regression); the two failure modes carry different clinical costs, and only the full precision-recall breakdown in Table 1, read alongside the confusion matrices in Figures 1 and 2, makes that difference visible.

For a screening application, where the clinical cost of a missed diagnosis (a false negative) is generally considered higher than the cost of a false alarm requiring a confirmatory test, recall is arguably the more clinically relevant metric of the two. On this basis, Logistic Regression's substantially higher recall (0.66 vs. 0.44) makes it the more defensible choice for a first-line screening tool in this study, despite its lower headline accuracy and identical F1-score to Random Forest.

At the same time, neither model is precise enough for standalone clinical deployment: precision remained low for both Logistic Regression (0.21) and Random Forest (0.25), meaning that roughly three to four of every five positive predictions were false alarms. This is consistent with the relatively small, six-variable feature set used here and suggests that additional clinical variables, such as HbA1c, family history detail, or lipid profile, and probability threshold calibration, would be necessary before such a model could be used beyond a coarse, high-recall pre-screening role that refers flagged patients for confirmatory laboratory testing rather than for diagnosis in itself.

4.5 Scope of the Evaluation and Its Limits

This study evaluates both models on a single dataset and a single 80/20 split, which bounds how far the comparative claims in Section 4.4 can be generalized. A more comprehensive evaluation would repeat the train-test split under k-fold cross-validation to report a mean and standard deviation for each metric rather than a single point estimate, and would test both models on at least one additional, independently collected dataset, such as the Pima Indians Diabetes dataset used throughout the related work in Section 2, or a clinical cohort from a different country, to establish whether the accuracy-recall trade-off observed here is a property of these two algorithms in general or an artifact of the DiaBD dataset's particular demographic composition and feature distribution. Neither extension was carried out for this paper, and the results reported here should be read as a single-dataset, single-split comparison rather than a claim of generalizable superiority for either model. We treat this as the paper's central limitation rather than a minor caveat, and it is the first item in the future-work agenda in Section 5.

4.6 Comparison with Prior Work

The accuracies obtained here (0.81-0.88) are lower than several PID-based studies reporting accuracies above 90-98% [5]-[10]. We attribute this gap primarily to two factors: first, the present study used a deliberately small, statistically justified feature set, six variables, rather than the full attribute set, prioritizing interpretability and avoiding overfitting to weakly predictive variables; and second, the DiaBD dataset is over six times larger and demographically broader than PID, which may make the classification task genuinely harder and the resulting metrics more representative of real-world performance than results obtained on a small, homogeneous, extensively mined benchmark dataset. This is consistent with the pattern noted by Deberneh and Kim [11] on a comparably large Korean cohort, where F1-scores (74%) were similarly more modest than typical PID-based results.

5. CONCLUSION AND FUTURE WORK

This study compared Logistic Regression and Random Forest for diabetes risk prediction on the DiaBD dataset using a statistically justified, six-variable feature set and a leakage-safe SMOTE correction for class imbalance. Random Forest achieved higher overall accuracy and AUC, while Logistic Regression achieved substantially higher recall for the diabetic class; the two models' F1-scores were statistically indistinguishable at 0.32. We conclude that no single model is unambiguously superior across all metrics, and that model selection for clinical screening tools should be driven by the relative cost of false negatives versus false positives in the intended deployment context, rather than by accuracy, AUC, or F1-score in isolation.

Future work should pursue five directions. First, and most directly in response to the scope limitation discussed in Section 4.5, both models should be re-evaluated under k-fold cross-validation and tested on at least one additional, independently collected dataset to determine whether the accuracy-recall trade-off reported here holds beyond the DiaBD dataset. Second, probability threshold tuning and cost-sensitive learning should be explored to improve the precision-recall balance of both models beyond the default 0.5 threshold used here. Third, additional clinically relevant variables, such as HbA1c, lipid profile, and dietary or physical activity indicators, should be incorporated to raise precision. Fourth, external validation on an independent population, such as a Nigerian hospital cohort, is needed to assess how well models trained on

the DiaBD dataset generalize outside Bangladesh, which is particularly relevant given that diabetes risk-factor prevalence varies across regions. Fifth, explainability techniques such as SHAP or LIME should be integrated so that clinicians can inspect the basis for individual predictions, which is essential for trust and adoption in real clinical workflows.

6. REFERENCES

[1] World Health Organization, "Global Health Observatory: Diabetes - prevalence and treatment, 1990-2022," WHO, Geneva, Switzerland, 2024. [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/761

[2] M. J. Hossain, M. Al-Mamun, and M. R. Islam, "Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused," Health Sci. Rep., vol. 7, no. 3, art. e2004, 2024.

[3] S. Safiri et al., "Prevalence, deaths and disability-adjusted life-years (DALYs) due to type 2 diabetes and its attributable risk factors in 204 countries and territories, 1990-2019," Front. Endocrinol., vol. 13, art. 838027, 2022.

[4] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432-439, 2021.

[5] 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.

[6] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," Comput. Methods Programs Biomed., vol. 220, art. 106773, 2022.

[7] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," BMC Bioinformatics, vol. 24, art. 224, 2023.

[8] A. Doğru, S. Buyrukoğlu, and M. Arı, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," Med. Biol. Eng. Comput., vol. 61, no. 3, pp. 785-797, 2023.

[9] T. Alghamdi, "Prediction of diabetes complications using computational intelligence techniques," Appl. Sci., vol. 13, no. 5, art. 3030, 2023.

[10] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning," IEEE Access, vol. 10, pp. 8529-8538, 2022.

[11] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," Int. J. Environ. Res. Public Health, vol. 18, no. 6, art. 3317, 2021.

[12] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," Healthc. Technol. Lett., vol. 10, no. 1-2, pp. 1-10, 2023.

[13] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," Comput. Methods Programs Biomed., vol. 202, art. 105968, 2021.

[14] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," Complexity, vol. 2021, art. 5525271, 2021.

[15] J. Zhao et al., "Attention-oriented CNN method for type 2 diabetes prediction," Appl. Sci., vol. 14, no. 10, art. 3989, 2024.

[16] T. T. Prama, M. J. Rahman, M. Zaman, F. Sarker, and K. A. Mamun, "DiaBD: A diabetes dataset for enhanced risk analysis and research in Bangladesh," Data Brief, art. 111746, 2025, doi: 10.1016/j.dib.2025.111746.

[17] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[18] 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.

[19] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[20] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[21] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: A systematic review," IEEE J. Biomed. Health Inform., vol. 25, no. 7, pp. 2744-2757, 2021.