

Early Disease Risk Prediction Using Health Records: A Comparative Study of Logistic Regression, Random Forest and XGBoost with SMOTE and SHAP Explainability

Anand Kale¹, Anisha Jadhav², Amrita Gupta³, Dr. Kavita Dhakad⁴

School of Information Technology, Indira University, Pune, Maharashtra, India

Abstract—Early identification of disease risk from routine clinical records can support timely assessment and preventive care. This paper presents a comparative study of three supervised machine learning classifiers, Logistic Regression, Random Forest and XGBoost, for predicting diabetes from structured health records. Experiments use the Pima Indians Diabetes dataset, which contains 768 records and eight clinical features. The pipeline treats physiologically implausible zero readings as missing values, applies median imputation and feature standardisation, uses a stratified 80:20 train-test split, and applies SMOTE to the training set only to reduce class imbalance. Models are evaluated on the same held-out test set using accuracy, precision, recall, F1-score and AUC-ROC, and SHAP is used to explain the XGBoost model. Random Forest and XGBoost both achieved the highest accuracy (0.747) and F1-score (0.649). All three models identified 36 of the 54 diabetic test cases (recall 0.667), while the tree-based models produced fewer false positives (21 versus 27) than Logistic Regression. Random Forest had the highest AUC-ROC (0.823). SHAP analysis identified Glucose, BMI and Age as the most influential features. The study is a research prototype and has not been clinically validated.

Keywords—Diabetes, disease risk prediction, explainable AI, machine learning, Random Forest, SHAP, SMOTE, XGBoost

I. INTRODUCTION

Early identification of disease risk is an important goal of preventive healthcare. Conditions such as diabetes develop gradually, and recognising risk earlier can support timely clinical assessment and lifestyle or medical intervention. Healthcare systems routinely record structured information such as demographic details, physiological measurements and laboratory values, and these records may contain patterns associated with disease outcomes.

Supervised machine learning can learn such patterns from historical records. However, three practical issues limit its usefulness. First, clinical datasets often contain missing or implausible values. Second, outcome classes are usually unequal, so a classifier may reach high accuracy while missing many positive cases. Third, a prediction is more useful when the factors behind it can be inspected.

This paper addresses these issues in a single, reproducible workflow. The contributions are: (1) a fair comparison of Logistic Regression, Random Forest and XGBoost under a common preprocessing pipeline and a common held-out test set; (2) leakage-aware handling of class imbalance by applying SMOTE to the training data only; (3) evaluation using precision, recall, F1-score and AUC-ROC in addition to accuracy; and (4) SHAP-based interpretation of the XGBoost model.

II. RELATED WORK

Machine learning has been widely studied for clinical prediction using structured patient data. Chen and Asch discuss both the opportunities and the limitations of machine learning in medicine [1], and Weng et al. show that machine learning can improve cardiovascular risk prediction from routine clinical data [2]. Logistic Regression remains a common interpretable baseline, while tree-based ensembles such as Random Forest and gradient boosting, including XGBoost [5], can capture nonlinear relationships and interactions.

Class imbalance is a recurring concern in medical datasets. Chawla et al. introduced SMOTE, which generates synthetic minority-class examples [3]. For a valid evaluation, resampling must be confined to the training data so that the test set remains an independent sample.

Interpretability is another active area. Lundberg and Lee proposed SHAP, which uses Shapley-value concepts to attribute a model output to individual features [4]. Recent diabetes prediction studies combine machine learning with SHAP, feature selection or class-imbalance handling [6]–[11]. In particular, studies on the Pima dataset and related clinical or survey data report competitive performance for tree-based models and use SHAP to highlight glucose-related and body-composition features [7]–[11].

The gap addressed here is practical: bringing together imbalance-aware training, a like-for-like comparison of several models, and feature-level explanation in one accessible workflow, while reporting limitations honestly.

III. METHODOLOGY

A. Dataset

The experiments use the Pima Indians Diabetes dataset [12], a widely used benchmark for binary diabetes classification. It contains 768 records, eight numerical input features and a binary Outcome label (0 = non-diabetic, 1 = diabetic). The features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. The dataset has 500 non-diabetic and 268 diabetic cases, an approximate 65%:35% split.

B. Preprocessing

Several physiological variables contain zero values that are not plausible as true measurements. These zero values are treated as missing and replaced by the median of the respective feature. Numerical features are standardised using z-score scaling. The data are divided into 80% training and 20% test subsets using stratified sampling with random seed 42, giving a test set of 154 records (100 non-diabetic and 54 diabetic).

C. Handling Class Imbalance

SMOTE is applied to the scaled training set only, so that the models are fitted on balanced classes while the test set remains untouched and reflects the original class distribution.

D. Models

Three classifiers are trained on the prepared training data:

1) *Logistic Regression*: models the log-odds of the outcome as a linear function of the features and serves as an interpretable baseline.

2) *Random Forest*: combines many decision trees trained on bootstrapped samples and random feature subsets, and can capture nonlinear patterns.

3) *XGBoost*: builds trees sequentially, with later trees correcting the errors of earlier ones and regularisation controlling complexity.

E. Evaluation Metrics

All models are evaluated on the same held-out test set using accuracy, precision, recall, F1-score, AUC-ROC and confusion matrices. In a screening-oriented task, recall is especially important because false negatives are diabetic cases that the classifier fails to identify; precision and AUC-ROC provide complementary information.

F. Explainability

SHAP TreeExplainer is applied to the trained XGBoost model to rank features by their contribution to the model output.

IV. RESULTS AND DISCUSSION

A. Exploratory Analysis

Fig. 1 shows the distribution of each feature by outcome class. Diabetic cases tend to have higher Glucose, BMI and Age values, with the clearest separation for Glucose.

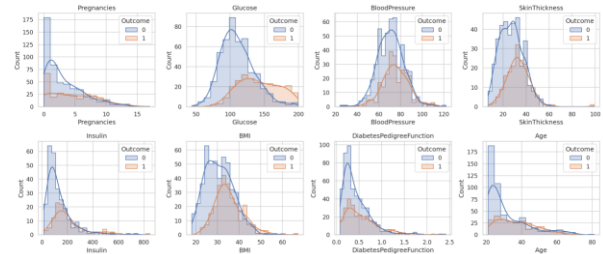


Fig. 1. Feature distributions by outcome class (0 = non-diabetic, 1 = diabetic).

The correlation matrix in Fig. 2 shows that Glucose has the strongest linear correlation with the Outcome label (0.49), followed by BMI (0.31) and Insulin (0.30); Skin Thickness (0.26) and Age (0.24) are weaker. Several predictors are also correlated with each other, notably BMI with Skin Thickness (0.65), Glucose with Insulin (0.58) and Age with Pregnancies (0.54). Correlation is descriptive and does not establish causation.

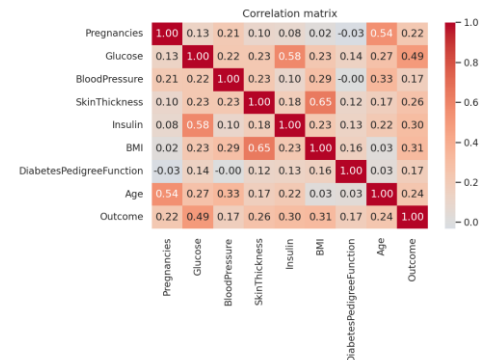


Fig. 2. Correlation matrix of the clinical features and the Outcome label.

B. Model Performance

Table I compares the three models on the held-out test set of 154 records.

TABLE I
PERFORMANCE COMPARISON ON THE HELD-OUT TEST SET

Metric	Logistic Regression	Random Forest	XGBoost
Accuracy	0.708	0.747	0.747
Precision	0.571	0.632	0.632
Recall	0.667	0.667	0.667
F1-score	0.615	0.649	0.649
AUC-ROC	0.810	0.823	0.809

Random Forest and XGBoost produced identical confusion matrices (Fig. 4) and therefore identical accuracy (0.747), precision (0.632), recall (0.667) and F1-score (0.649). Both outperformed Logistic Regression on accuracy (0.708), precision (0.571) and F1-score (0.615). The recall of all three models is the same. The models differ in AUC-ROC: Random Forest is highest (0.823), followed by Logistic Regression (0.810) and XGBoost (0.809). The ROC curves in Fig. 3 overlap closely, so these differences are small.

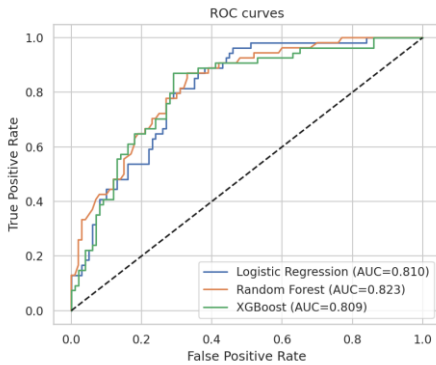


Fig. 3. ROC curves and AUC-ROC values for the three models.

C. Class-wise Results

For the positive (diabetic) class, with a support of 54 test cases, each model correctly identified 36 diabetic cases and missed 18. The difference lies in false positives: Logistic Regression wrongly flagged 27 of the 100 non-diabetic cases, while Random Forest and XGBoost flagged 21 (Fig. 4). Table II summarises the class-1 metrics.

TABLE II
CLASS-1 (DIABETIC) RESULTS ON THE TEST SET

Model	Precision	Recall	F1-score	Support
Logistic Regression	0.57	0.67	0.62	54
Random Forest	0.63	0.67	0.65	54
XGBoost	0.63	0.67	0.65	54

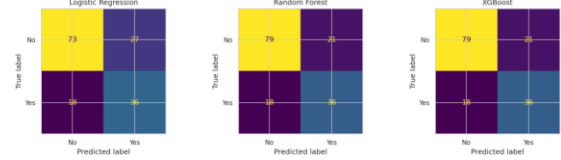


Fig. 4. Confusion matrices of the three models on the test set.

D. Explainability

The SHAP summary plot for XGBoost (Fig. 5) ranks Glucose as the most influential feature, followed by BMI, Age and Diabetes Pedigree Function. Higher Glucose and BMI values push the model output toward the diabetic class. Insulin, Pregnancies, Blood Pressure and Skin Thickness contribute comparatively little in the fitted model. Insulin correlates with the Outcome label about as strongly as BMI (0.30 versus 0.31) but receives much less weight in SHAP, probably because it overlaps with Glucose (correlation 0.58). SHAP values describe how the trained model uses its inputs; they do not prove causal relationships.

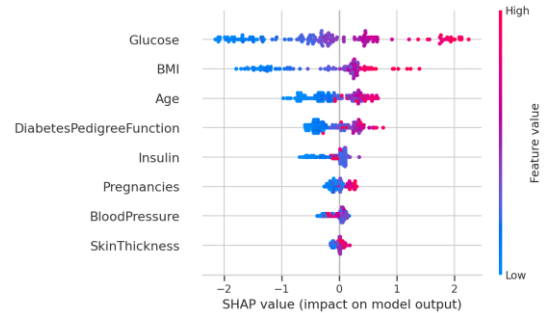


Fig. 5. SHAP summary plot for the XGBoost model.

E. Discussion

The differences between models are modest, which is expected on a small dataset. All three models miss the same 18 diabetic cases, so none is better at finding diabetic patients at the default decision threshold. Random Forest and XGBoost reduce false positives from 27 to 21, which raises precision and accuracy over Logistic Regression. Because the AUC-ROC values differ by only 0.014 on a single test set of 154 records, the models should be regarded as performing similarly, and the choice between them should follow the intended screening use, interpretability needs and threshold selection rather than a single metric.

V. LIMITATIONS

- The dataset has only 768 records and may not represent other populations or care settings.
- Results come from a single stratified train-test split and may vary with a different split; cross-validation was not used for the reported results.

- Every model missed 18 of the 54 diabetic test cases (33%), so the models are not suitable as a stand-alone screening tool.
- Median imputation values were computed on the full dataset before the train-test split, which may allow minor information from the test set to influence preprocessing. Future work will fit the imputer on the training data only.
- SMOTE creates synthetic training examples and does not replace validation on real, independent clinical data.
- SHAP analysis was performed for XGBoost only, and explanation stability across patients and models was not studied.
- The system is a research prototype and is not a diagnostic tool.

VI. CONCLUSION AND FUTURE WORK

This paper compared Logistic Regression, Random Forest and XGBoost for diabetes risk prediction from structured health records, using median imputation, standardisation, training-set SMOTE and SHAP-based explanation. Random Forest and XGBoost gave the same accuracy (0.747) and F1-score (0.649), all three models found 36 of 54 diabetic cases, and Random Forest had the highest AUC-ROC (0.823). Glucose, BMI and Age were the most influential features in the SHAP analysis. The findings show that model differences on this benchmark are small, and that the results are those of an academic prototype rather than clinical evidence.

Future work includes stratified k-fold cross-validation, probability calibration and decision thresholds chosen to reduce false negatives, evaluation on larger and more diverse independent datasets, assessment of SHAP explanation stability, and development of a simple screening interface with external and prospective clinical validation before any real-world use.

Acknowledgment

The authors thank the School of Information Technology, Indira University, Pune, for providing the academic environment for this research.

REFERENCES

- [1] J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine – Beyond the Peak of Inflated Expectations," *New England Journal of Medicine*, 2017.
- [2] S. F. Weng, J. Reys, J. Kai, J. M. Garibaldi and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLOS ONE*, 2017.
- [3] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [6] X. Chai, "Research on diabetes risk prediction based on machine learning," *Proceedings of the 2025 International Conference on Health Informatization and Data Analytics*, 2025. doi: 10.1145/3759972.3760174.
- [7] L. Wang, X. Wang, A. Chen, X. Jin and H. Che, "Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model," *Healthcare*, vol. 8, no. 3, p. 247, 2020. doi: 10.3390/healthcare8030247.
- [8] Q. Liu et al., "Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques," *Journal of Personalized Medicine*, vol. 12, no. 6, p. 905, 2022. doi: 10.3390/jpm12060905.
- [9] I. Kırbaş and A. Çifci, "Leveraging SHAP for interpretable diabetes prediction: A study of machine learning models on the Pima Indians Diabetes dataset," *Balkan Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 128–139, 2025. doi: 10.17694/bajece.1577929.
- [10] P. K. Mohanty et al., "Leveraging Shapley additive explanations for feature selection in ensemble models for diabetes prediction," *Bioengineering*, vol. 11, no. 12, p. 1215, 2024. doi: 10.3390/bioengineering11121215.
- [11] P. Netayawijit, W. Chansanam and K. Som-In, "Interpretable machine learning framework for diabetes prediction: Integrating SMOTE balancing with SHAP explainability for clinical decision support," *Healthcare*, vol. 13, no. 20, p. 2588, 2025. doi: 10.3390/healthcare13202588.
- [12] Kaggle, "Pima Indians Diabetes Database." Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>