Scientific Research Publishing

Journal of Software Engineering and Applications, 2026, 19(2), 7-24
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

A Hybrid Machine Learning Framework for Early Diabetes Prediction in Sierra Leone Using Feature Selection and Soft-Voting Ensemble

Aminata Bah

College of Software, Nankai University, Tianjin, China
Email: amienabah133@gmail.com

How to cite this paper: Bah, A. (2026) A Hybrid Machine Learning Framework for Early Diabetes Prediction in Sierra Leone Using Feature Selection and Soft-Voting Ensemble. *Journal of Software Engineering and Applications*, **19**, 7-24.
https://doi.org/10.4236/jsea.2026.192002

Received: January 18, 2026
Accepted: February 22, 2026
Published: February 25, 2026

Abstract

This paper proposes a hybrid machine learning framework for early diabetes prediction tailored to Sierra Leone, where locally representative datasets are scarce. The framework integrates Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) into a probability-based soft-voting ensemble that prioritizes sensitivity (recall) for screening. Experiments were conducted under two conditions: 1) using all available features and 2) after feature selection based on RF importance, retaining six clinically meaningful predictors (Glucose, Diabetes Pedigree Function, Skin Thickness, Age, Body Mass Index, and Insulin). Evaluation employed Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion-matrix analysis with a screening-oriented decision threshold. Before feature selection, the hybrid model achieved a Recall of 0.8571 and an ROC-AUC of 0.8610, reducing false negatives compared with individual classifiers. After feature selection, performance remained competitive while improving interpretability and deployment feasibility. Benchmark validation on the Pima Indians Diabetes dataset further supported the robustness of the approach. The proposed hybrid framework provides a practical, sensitivity-focused decision-support tool for early diabetes screening in low-resource clinical environments.

Keywords

Diabetes Prediction, Hybrid Ensemble, Soft Voting, Feature Selection, XGBoost, Sierra Leone

1. Introduction

Diabetes mellitus is a major global health burden, with prevalence projected to

rise in low- and middle-income countries as urbanization, dietary patterns, and more sedentary lifestyles take hold [1] [2]. In the African context, limited diagnostic facilities, the absence of screening programs, and the resulting delay in diagnosis mean that many patients are diagnosed only after complications from the condition have developed [2] [3]. In Sierra Leone, the situation is further worsened by limited resources, limited access to the laboratory, and an unreliable health information system [3] [4]. Machine learning (ML) techniques show strong potential for identifying risks by learning indicators that predict the likelihood of diabetes at an early stage [5] [6]. However, implementing ML models in resource-limited settings is not straightforward. They may not perform well because of imbalance, missingness, small datasets, and the use of indicators that require laboratory analysis, which is not always accessible at primary healthcare units [7] [8]. In addition, most existing models prioritize accuracy without recognizing that this may not always yield good results for imbalanced datasets, such as those in medical domains, where recall is more important [7]. This paper presents a hybrid soft-voting ensemble of Random Forest, Logistic Regression, and XGBoost. The model focuses on developing a high-recall classification strategy because missed diabetes diagnoses can lead to complications from delayed treatment [3] [7]. The second aim of this study is to assess the model's practicality by testing it both before and after feature selection. Feature selection enhances the model's interpretability for practical use in Sierra Leone.

2. Literature Review

2.1. Overview of Diabetes and Early Detection

Diabetes mellitus remains a major global public health burden, with prevalence projected to rise substantially in low- and middle-income countries [1] [2]. Early detection is essential to reduce complications and premature mortality; however, limited diagnostic capacity and weak routine screening constrain early diagnosis in many low-resource settings [3] [4].

2.2. Machine Learning Models for Diabetes Prediction

Machine learning has been widely applied to diabetes prediction using clinical and demographic indicators. Logistic Regression is commonly used due to interpretability, Random Forest for robust nonlinear modeling, and XGBoost for strong discrimination via regularized boosting [9]-[11]. In screening contexts, recall and ROC-AUC are often emphasized to reduce missed diabetic cases under class imbalance [7] [12].

2.3. Feature Selection, Interpretability, and Feasibility in Clinics

Feature feasibility is critical in low-resource clinics where laboratory-intensive biomarkers may be unavailable. Feature selection (e.g., Random Forest importance) can improve interpretability and reduce data-collection burden while maintaining screening performance [7] [10].

2.4. Research Gap and Motivation for This Study

There is no publicly available diabetes dataset specific to Sierra Leone, and many prior models are trained on non-African datasets with different population characteristics [1] [4]. This study addresses these gaps by developing a Sierra Leone-oriented dataset, prioritizing clinically feasible predictors, and evaluating a hybrid soft-voting ensemble with benchmark validation on the Pima dataset [13] [14].

3. Methodology

This section discusses the overall pipeline approach for diabetes prediction modeling, from data preparation to model training and evaluation. This research uses a synthetic Sierra Leone-focused dataset for comparison with the Pima Indians dataset, performs a stratified train-test split, addresses class imbalance, and applies data preprocessing prior to training models using Logistic Regression, Random Forest, and XGBoost algorithms. A hybrid soft-voting approach and Random Forest variable selection are further employed, along with overall screening-relevant model evaluation criteria. Figure 1 illustrates the overall flow process for the proposed approach.

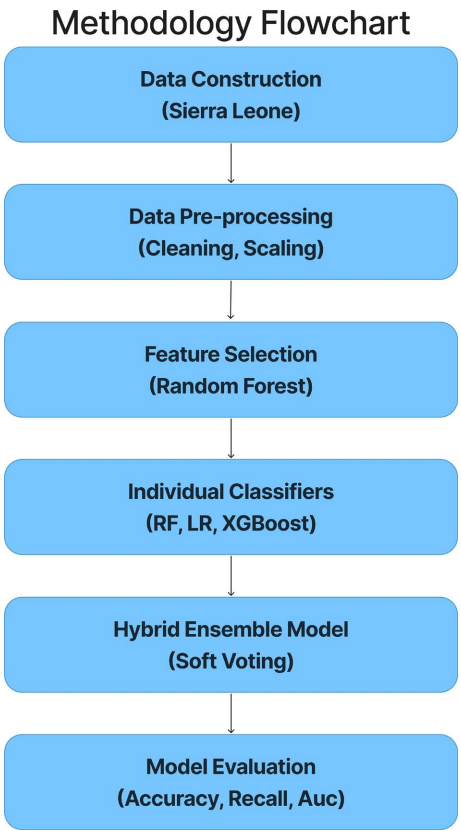


Figure 1. Methodology workflow diagram.

As illustrated in Figure 1, the data construction process for model evaluation is streamlined. The figure encapsulates the entire process, beginning with the crea-

tion of the dataset for Sierra Leone, moving on to data preprocessing, Random Forest-based feature selection, training of separate models for classification using RF, LR, and XGBoost, and finally generating predictions for the models using a soft voting hybrid technique, with model accuracy determined by metrics of relevance for screening.

3.1. Dataset Description

3.1.1. Sierra Leone Synthetic Diabetes Dataset

This study used a Sierra Leone-oriented diabetes dataset designed to reflect plausible demographic and clinical characteristics relevant to local screening contexts. The dataset contains $N = 600$ adult records with nine predictors (Sex, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age) and a binary outcome label (Outcome), where 1 indicates diabetic and 0 indicates non-diabetic. The class distribution is imbalanced, with 103 diabetic cases (17.17%) and 497 non-diabetic cases (82.83%), supporting the need for sensitivity-focused evaluation in screening. Because the dataset is synthetic, it contains no identifiable patient information while enabling controlled experimentation in data-scarce settings such as Sierra Leone. The Sierra Leone-oriented dataset was generated using probabilistic sampling guided by epidemiological targets and clinically plausible ranges reported in public health sources. Parameter ranges and distributions were chosen to be consistent with reported diabetes prevalence and commonly used clinical screening variables, then values were sampled and constrained to realistic physiological bounds. The outcome label distribution was calibrated to reflect the intended prevalence within the constructed dataset while preserving plausible relationships among core predictors used in diabetes risk assessment. This approach provides a privacy-preserving dataset for experimentation in a data-scarce setting, while acknowledging that synthetic data may not fully capture real-world clinical variability [2] [4].

3.1.2. Pima Indians Diabetes Dataset (Benchmark Validation)

To validate the models' generalization capabilities on a dataset other than the Sierra Leone Synthetic Dataset, the Pima Indians Diabetes Dataset was employed. This dataset is publicly available through the UCI Machine Learning Repository and has been widely used in machine learning studies as a standard benchmark for evaluating diabetes classification algorithms [15]. The benchmark test on a public dataset helps ensure reproducibility and also enables comparison with past approaches, as it provides further evidence that the proposed modeling pipeline still performs well beyond the built dataset. Nonetheless, since the Pima dataset pertains to a specific population, it cannot serve as a substitute for clinical validation in Sierra Leone, and it should be tested on real data when available [13] [15].

3.2. Experimental Design and Data Splitting

3.2.1. Train-Test Split

The data is split using an 80:20 stratified train-test split to maintain the class pro-

portion of diabetic and non-diabetic samples. A fixed random seed (random_state = 42) was used to ensure reproducibility and enable consistent comparisons across models. Stratification is important, especially when working with imbalanced datasets, to maintain unbiased evaluation metrics [12].

3.2.2. Handling Class Imbalance

To reduce class imbalance and improve diabetic case detection, class weighting was applied to the Logistic Regression and Random Forest classification algorithms. In XGBoost, class imbalance was handled by setting the scale_pos_weight parameter to the negative-to-positive instances ratio in the training data set. Such an approach enables sensitivity-centric learning for classification models and prevents underdetection of classes during screening assessments [7] [12] [16].

3.3. Data Preprocessing

Prior to model training, missing values for continuous variables were replaced with the median. To complete the benchmark validation task for the Pima dataset, medically implausible values of zero in physiological variables such as glucose level, blood pressure, skin thickness, insulin level, and BMI were treated as missing values. Missing values were imputed with appropriate values. The Sex attribute was transformed into an indicator variable. Continuous predictor variables were scaled for logistic regression modeling because scaling enables more stable optimization and comparable scaling of model coefficients. However, tree-based models such as Random Forest and XGBoost were trained on the predictor variables without scaling, as they are scale-invariant.

3.4. Models and Training Configuration

3.4.1. Models

Three baseline models were trained:

- Random Forest (RF): A bagging-based ensemble of decision trees that improves generalization and robustness [10].
- Logistic Regression (LR): A linear probabilistic classifier widely used in clinical prediction for its interpretability [9].
- XGBoost: A regularized gradient boosting method that models nonlinearities efficiently and often achieves strong ROC-AUC on structured data [11].

The experiments were conducted in Python, employing scikit-learn to build and test the models and XGBoost for the gradient boosting algorithm. The dataset was split 80:20 using stratified sampling with a fixed random state (random_state = 42). The class imbalance was handled through class-weighting for Logistic Regression and Random Forest, and through the scale_pos_weight parameter, which was computed from the class ratio in the data, for XGBoost.

3.4.2. Base Learners and Hyperparameter Settings

Unless otherwise stated, the models were trained with the following key hyperparameters:

- Logistic Regression (LR): max_iter = 3000, class_weight = "balanced", random_state = 42.
- Random Forest (RF): n_estimators = 400, max_depth = None, class_weight = "balanced", random_state = 42.
- XGBoost (XGB): n_estimators = 500, learning_rate = 0.05, max_depth = 4, subsample = 0.9, colsample_bytree = 0.9, reg_lambda = 1.0, eval_metric = "logloss", random_state = 42, and scale_pos_weight = (N_negative/N_positive).

3.4.3. Software Environment

All experiments were implemented in Python 3.13.5 (Anaconda distribution) using scikit-learn 1.6.1 for model development and evaluation, and XGBoost 3.1.2 for gradient boosting [17].

3.5. Hybrid Soft-Voting Ensemble

To leverage the complementary strengths of the three base learners, a hybrid soft-voting ensemble was constructed using Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). Each base model outputs a predicted probability of the positive class (diabetes), denoted as $P_{RF}(y=1|x)$, $P_{LR}(y=1|x)$, and $P_{XGB}(y=1|x)$. The hybrid ensemble computes the final probability by averaging these outputs:

$$P_{\text{hybrid}}(y=1|x)=\frac{P_{RF}(y=1|x)+P_{LR}(y=1|x)+P_{XGB}(y=1|x)}{3}. \quad (1)$$

The hybrid ensemble architecture follows a probability-based soft-voting design. After preprocessing, the three base learners (Random Forest, Logistic Regression, and XGBoost) are trained, and each outputs a predicted probability of diabetes for a given sample. The final hybrid probability is computed as the average of the three predicted probabilities (soft voting). A screening-oriented decision threshold of 0.3 is then applied to convert probabilities into class labels, where samples with $P_{\text{hybrid}}(y=1|x)\geq 0.3$ are classified as diabetic (1), and otherwise as non-diabetic (0). This design prioritizes recall to reduce false negatives, which is clinically preferable in early screening settings where missed cases carry a higher risk than false positives.

The decision threshold of 0.3 was selected as a screening-oriented operating point to prioritize recall and reduce missed diabetic cases. Future work will formalize threshold selection using precision-recall analysis or F_β optimization (with $\beta>1$) to select an operating point aligned with clinical screening objectives.

3.6. Feature Selection Strategy (Top-6 Features)

To select features, the Random Forest feature importance ranking was used, in which features were ranked by their importance in reducing impurities in the trees. This technique reduces the number of features and enhances the interpretation of the results by selecting the most relevant features [10] [18]. The first six

features selected were Glucose, Diabetes Pedigree Function (DPF), Skin Thickness, Age, Body Mass Index (BMI), and Insulin. These features are clinically relevant indicators of diabetes risk and are commonly used in diabetes screening research, making the resulting model more interpretable and feasible for low-resource settings. Feature selection was performed only on the training split, and the selected feature subset was then applied to the test split, ensuring that no test-set information influenced feature ranking and preventing data leakage.

Although the selected features are commonly used in diabetes prediction research, availability may vary across facilities; therefore, future deployment can use a minimal subset (e.g., glucose, BMI, age) in settings where insulin or detailed family-history measures are not routinely obtainable [3] [4] [8] [19].

3.7. Evaluation Metrics and Validation Strategy

For evaluating the models' performance, Accuracy, Precision, Recall, Sensitivity, Specificity, F1-score, and ROC-AUC were considered. Moreover, the confusion matrix was analyzed to assess the measures of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), providing a detailed interpretation of classification errors that is important for screening-oriented medical applications. Because the datasets are imbalanced; hence, accuracy may not provide a fair measure of the models' performance. Sensitivity- and threshold-independent metrics, such as ROC-AUC, were considered to better evaluate the models' performance [7] [12]. Benchmark validation was conducted using the Pima Indians Diabetes Dataset as an external public dataset to provide additional evidence of model robustness beyond the Sierra Leone synthetic dataset [13] [15] [20].

4. Results

4.1. Results before Feature Selection (All Features)

Table 1. Performance metrics before feature selection.

Model	Accuracy	Precision	Recall	F1	ROC_AUC
RF (All)	0.783333	0.432432	0.761905	0.551724	0.869649
LR (All)	0.616667	0.288136	0.809524	0.425000	0.815296
XGB (All)	0.766667	0.410256	0.761905	0.533333	0.840789
Hybrid (All)	0.775000	0.428571	0.857143	0.571429	0.860991

Table 1 provides a summary of classification performance on the three individual models and the hybrid ensemble with all predictors considered. Since screening should attempt to miss as few patients with diabetes as possible, recall and confusion matrix error profiling are considered more important than accuracy measures per se. The hybrid model demonstrates improved sensitivity relative to most of the individual models, suggesting that probability combination may be a

useful approach in detecting at-risk patients with a fair level of discriminability, as indicated by ROC-AUC values. Based on this analysis, RF had an accuracy of 0.7833 and an ROC-AUC of 0.8696, indicating very strong discriminability, although LR had a relatively high recall of 0.8095 with a lower precision of 0.2881, indicating a tendency toward more false positives. XGBoost had a relatively stable performance with an ROC-AUC of 0.8408. The hybrid model had the highest recall of 0.8571 with a very strong ROC-AUC of 0.8610, further indicating that combining different models with complementary strengths may be useful [21] [22].

4.1.1. Confusion Matrix Analysis

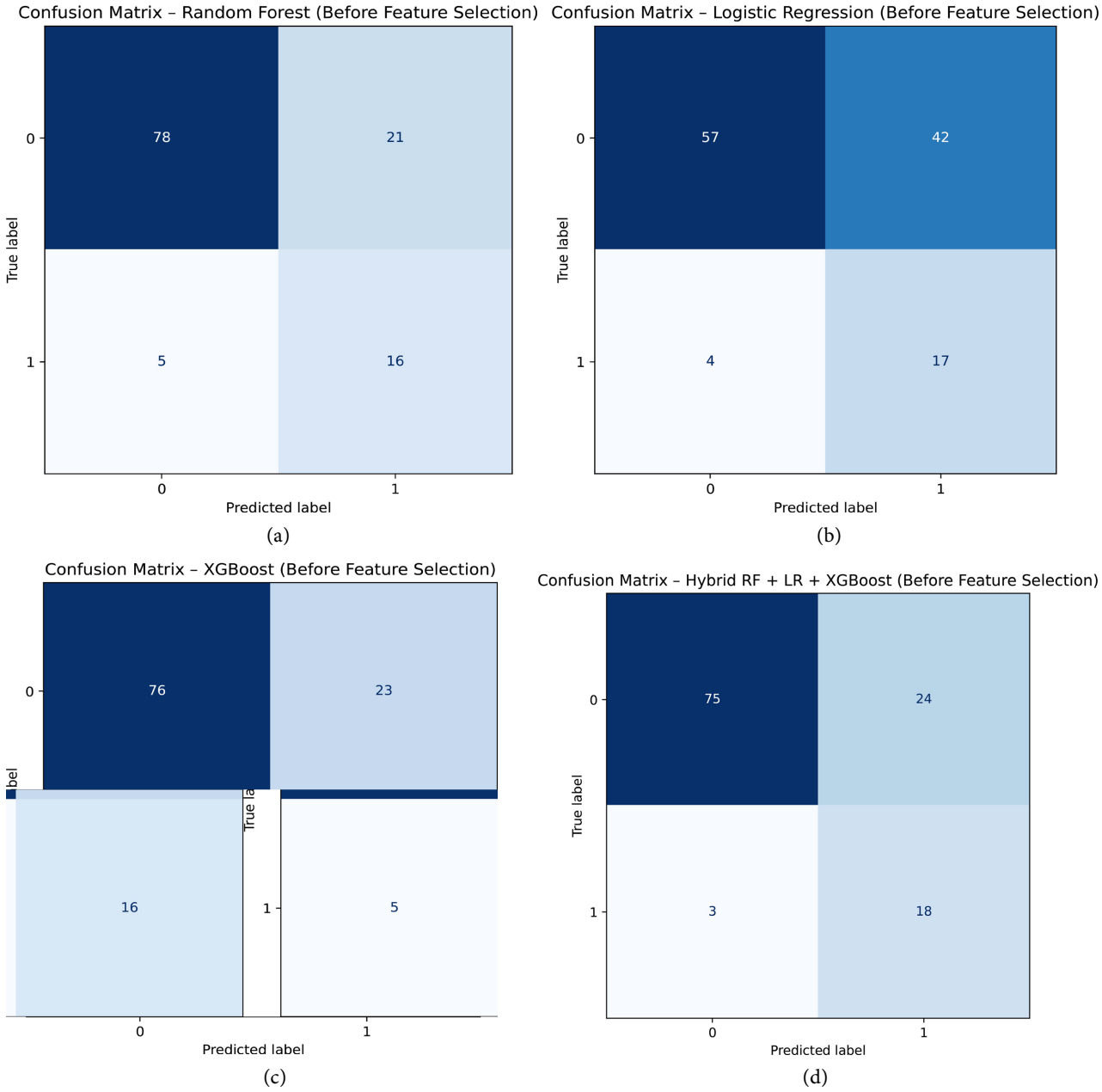


Figure 2. Confusion matrices before feature selection (all features) at the 0.3 screening threshold.

The confusion matrices in Figure 2 show the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the 0.3 threshold. LR had fewer false negatives than other individual models but a large number of false positives, indicating a sensitivity-driven boundary. RF and XGBoost handled false positives better but missed more diabetics than other models. The hybrid model offered a better tradeoff, reducing false negatives while keeping false positives under better control than LR. For screening, while false positives can be identified through further testing, false negatives remain at risk of developing complications because they remain unidentified [3] [12].

4.1.2. ROC Curves/AUC Comparison

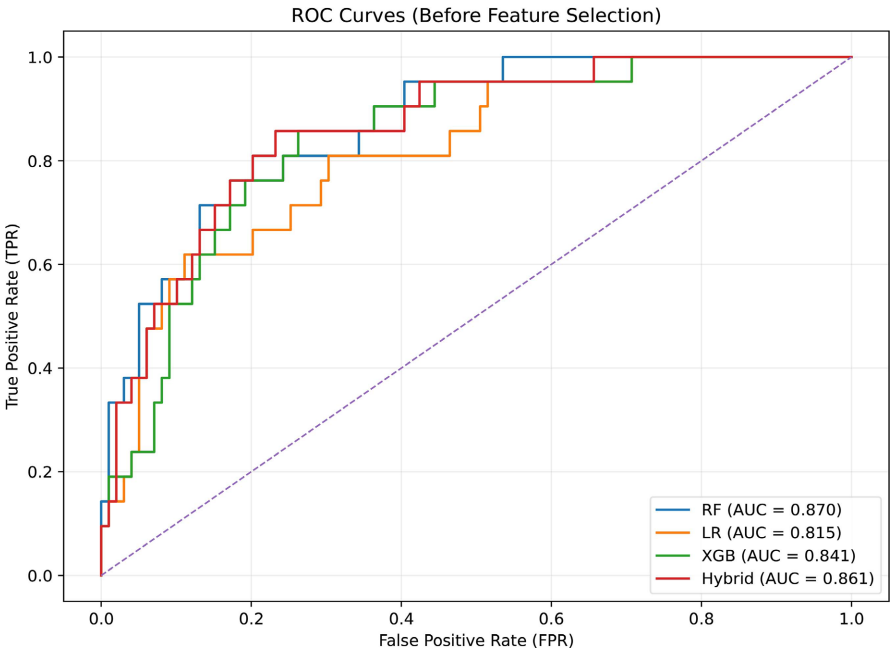


Figure 3. ROC curves (all features/before feature selection).

The ROC curves for the models before feature selection are shown in Figure 3. The ROC curves assess discrimination across all possible thresholds. Models with curves closer to the top-left corner and higher AUC values perform better at separating the diabetic and non-diabetic groups. The ROC-AUC index is threshold-independent, making it a useful supplement to the analysis of the confusion matrix at the screening operating point to ensure that any improvement in sensitivity does not come at the expense of discrimination quality. The highest AUC value (≈0.870) was obtained by the RF model, with the hybrid model second (≈0.861), suggesting that the aggregation operation in the ensemble model enhanced sensitivity without compromising discriminative ability [11] [21].

4.2. Results after Feature Selection (Top-6 Features)

4.2.1. Selected Features Summary

The chosen predictors, Glucose, DPF, Skin Thickness, Age, BMI, and Insulin, are

important factors in diabetes risk and remain practicable within the primary screening environment, with simple capabilities [3] [4] [8]. Reducing the features will also make it easier to interpret the results, enabling healthcare workers to understand which inputs drive risk predictions [9] [10].

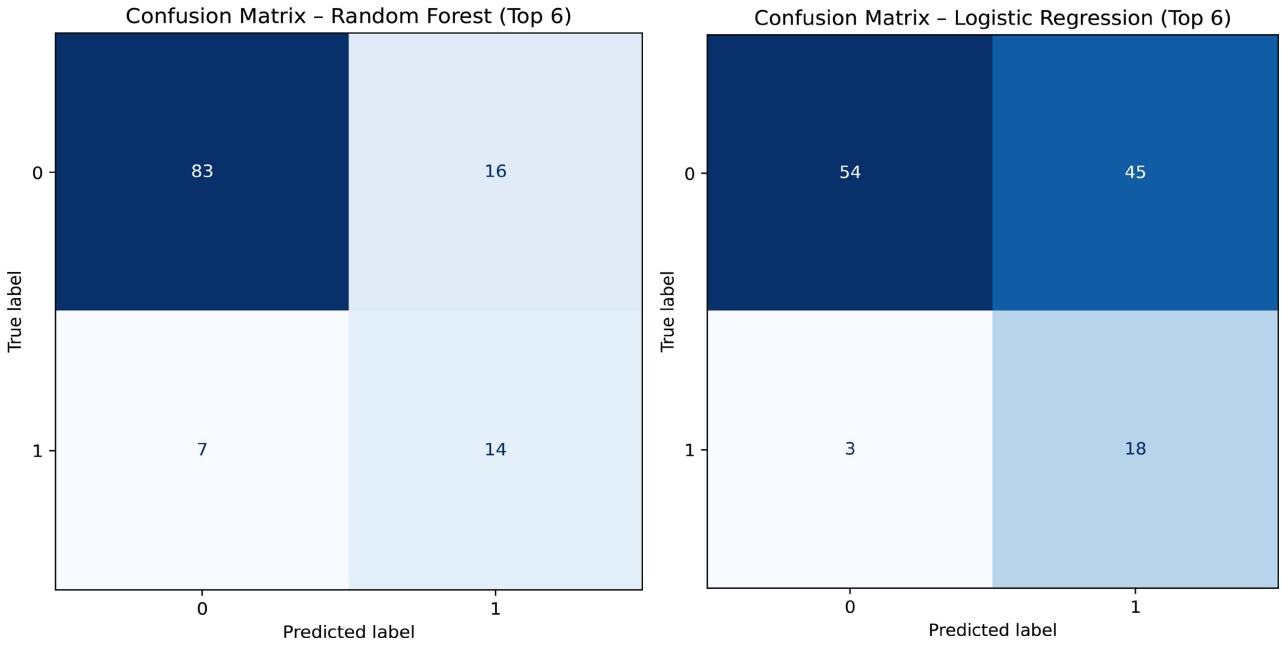
4.2.2. Performance Metrics (Top-6 Features)

Table 2. Performance metrics after feature selection (top-6 features).

Model	Accuracy	Precision	Recall	F1	ROC_AUC
RF (Top 6)	0.808333	0.466667	0.666667	0.549020	0.859548
LR (Top 6)	0.600000	0.285714	0.857143	0.428571	0.817701
XGB (Top 6)	0.808333	0.458333	0.523810	0.488889	0.839827
Hybrid (Top 6)	0.800000	0.457143	0.761905	0.571429	0.850409

After compressing the predictors into six interpretable features, as presented in Table 2, the results demonstrate competitive performance while improving interpretability and deployability. This is because the six predictors are easily available and measurable in many primary health settings. Changes in sensitivity at a constant threshold depend on the usual trade-off between interpretability and the point of operation in the screening problem. LR maintained the highest recall of 0.8571 but the lowest precision because of the high rate of false positives. RF and XGB achieved higher accuracy and precision than LR. This is because the ensemble method achieved a good trade-off between the two measures, maintaining a high recall of 0.7619 while improving precision compared to LR, and is therefore appropriate for screening and confirmatory testing [12] [21].

4.2.3. Confusion Matrices (Top-6 Features)



(a) (b)

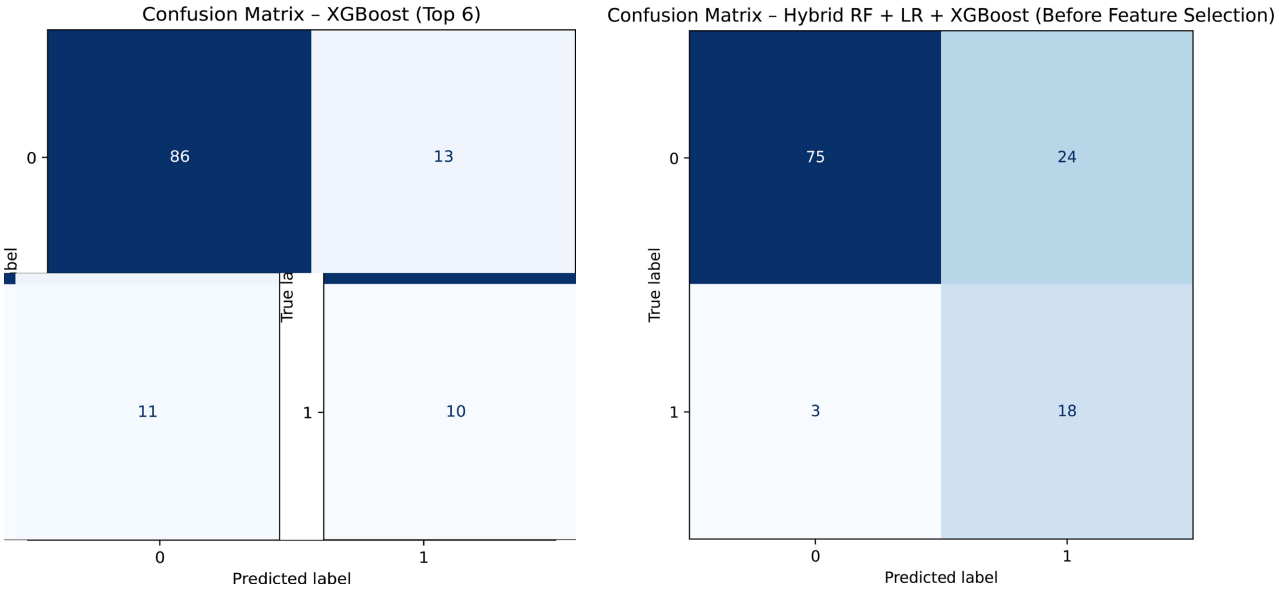


Figure 4. Confusion matrices after feature selection (top-6 features) at the 0.3 screening threshold.

Figure 4 shows the confusion matrices of the top 6 features after feature selection. LR had the lowest missed-diabetic rate (low FN) and a high FP rate, while RF and XGBoost had a lower FP rate at the expense of a higher missed-diabetic rate. The hybrid approach yielded a fair distribution of errors, which is preferred in screening, where missed cases are more harmful than false positives that can be tested again [3] [12].

4.2.4. ROC Curves/AUC Comparison (Top-6 Features)

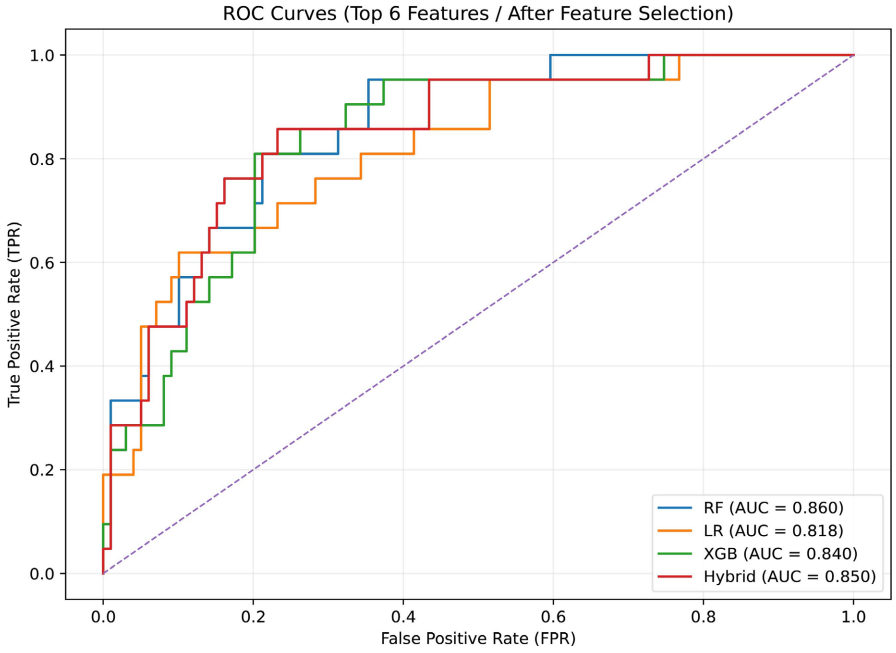


Figure 5. ROC curves (top-6 features/after feature selection).

As shown in Figure 5, when the number of features was reduced to six, the ROC-AUC scores stabilized, indicating that the compact set of features still contains substantial discriminative information. This supports the notion that the model is feasible for implementation, as fewer features can be used to preserve discriminative power while facilitating interpretation and data acquisition [7] [10].

4.3. Comparative Analysis: Before vs After Feature Selection

Table 3. Comparative performance before vs after feature selection.

Stage	ModelShort	Accuracy	Precision	Recall	F1	ROC_AUC
Before FS (All)	RF	0.783333	0.432432	0.761905	0.551724	0.869649
Before FS (All)	LR	0.616667	0.288136	0.809524	0.425000	0.815296
Before FS (All)	XGB	0.766667	0.410256	0.761905	0.533333	0.840789
Before FS (All)	Hybrid	0.775000	0.428571	0.857143	0.571429	0.860991
After FS (Top 6)	RF	0.808333	0.466667	0.666667	0.549020	0.859548
After FS (Top 6)	LR	0.600000	0.285714	0.857143	0.428571	0.817701
After FS (Top 6)	XGB	0.808333	0.458333	0.523810	0.488889	0.839827
After FS (Top 6)	Hybrid	0.800000	0.457143	0.761905	0.571429	0.850409

Comparison 3 shows how each model's performance changes with feature selection. Changes in precision and stability indicate a reduction in overfitting, while changes in recall underscore the importance of threshold values when a clinical application prioritizes high sensitivity. This comparison reinforces that feature selection should be evaluated not only by accuracy but also by its impact on false negatives.

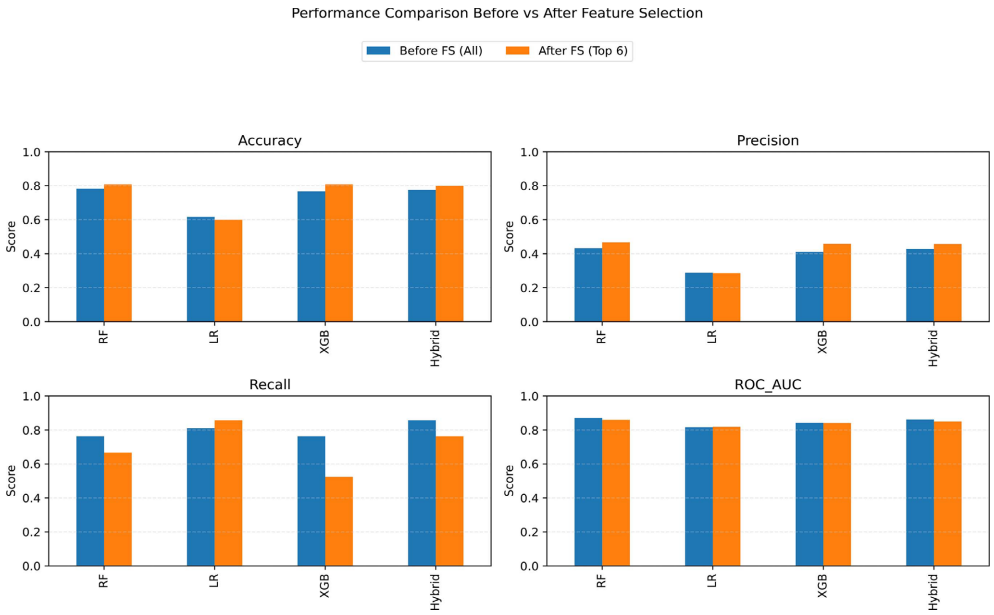


Figure 6. Performance comparison before vs after feature selection (accuracy, precision, recall, ROC-AUC).

Table 3 and Figure 6 highlight the performance of models before and after applying feature selection on each model type. Applying feature selection to the models increased their accuracy and precision for RF and XGBoost, and maintained stable AUC values across all models. The recall values of some models decreased at a certain threshold, consistent with a common screening result that reducing features can cause a slight change in the probability distribution and sensitivity at that threshold [7] [12]. This underscores the need for threshold selection and confusion-matrix interpretation when the clinical goal is high recall.

4.4. Benchmark Validation on the Pima Indians Diabetes Dataset

To ensure external comparability of the results, the experiments were performed on the Pima Indians Diabetes dataset, which is commonly used to predict diabetes [13] [14]. The preprocessing step treated impossible zeros in the physiological variables as missing data and imputed them with the median, as commonly practiced in previous studies on this dataset [13] [15].

Table 4. Benchmark results on Pima Indians diabetes dataset.

Model	Acc.	Bal. Acc.	Prec.	Rec.	Spec.	F1	AUC	TN	FP	FN	TP
Random Forest	0.7532	0.7120	0.6739	0.5741	0.8500	0.6200	0.8205	85	15	23	31
XGBoost	0.7403	0.7148	0.6296	0.6296	0.8000	0.6296	0.8172	80	20	20	34
Logistic Regression	0.7338	0.7269	0.6032	0.7037	0.7500	0.6496	0.8126	75	25	16	38
Hybrid (Soft Voting)	0.7338	0.7098	0.6182	0.6296	0.7900	0.6239	0.8256	79	21	20	34

A comparison of the Pima Indians dataset is provided in Table 4, which facilitates comparison with previous work. Variations are expected when comparing with the Sierra Leone-centric dataset due to differences in populations, prevalence, and measurement distributions. However, being competitive on ROC-AUC and Sensitivity/Specificity demonstrates that the modeling framework remains strong regardless of the dataset used [13].

The confusion matrix for the hybrid model (RF + LR + XGBoost) designed for the Pima Indians Diabetes dataset is shown in Figure 7, which depicts the classification distribution for True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The benchmark test results serve as an external assessment for comparison with the popular global dataset, thereby aiding the evaluation of the applicability of the hybrid model framework to other sources beyond the Sierra Leone-specific dataset. In the context of the study, the FN continues to be the most serious error from a medical viewpoint since it corresponds to the overlooked diabetic patients, whereas the FP corresponds to the people mistakenly diagnosed as having diabetes, thus needing confirmation tests.

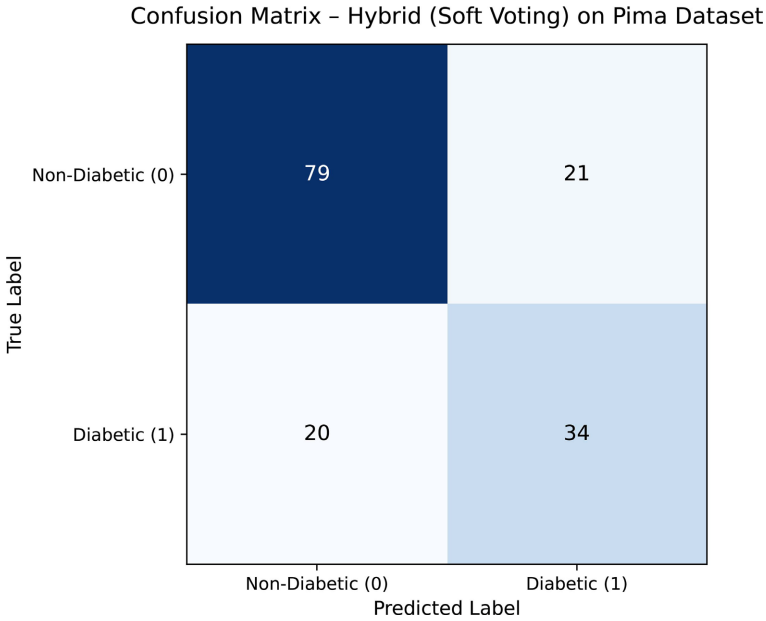


Figure 7. Confusion matrix—hybrid model on Pima dataset.

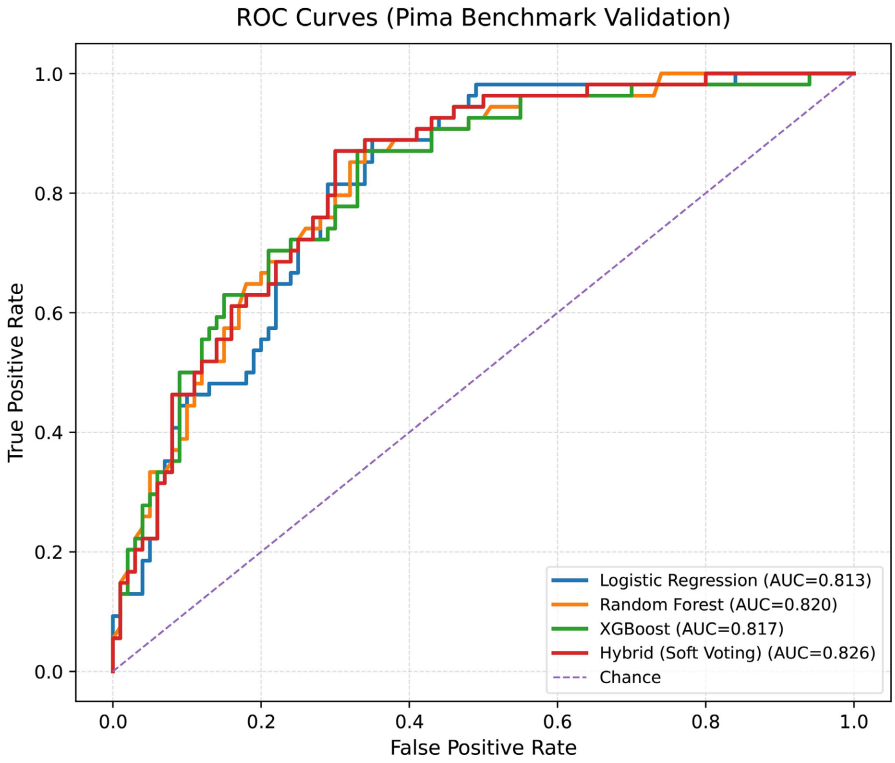


Figure 8. ROC curves—benchmark validation on Pima dataset.

As shown in Figure 8, the ROC curves for the benchmark validation on the Pima dataset illustrate the discrimination performance of the proposed framework across decision thresholds.

The benchmark results also show that the hybrid model maintained a competitive level of discrimination, with AUC values of approximately (≈0.8256). The

ROC curves illustrate the level of discrimination on a commonly used external dataset. There are also variations in performance on the Sierra Leone dataset due to population factors [13] [14].

5. Discussion

5.1. Model Behavior and Trade-Offs

The experiments highlight differences in classifier behavior across learning algorithms. RF performed well in distinguishing non-diabetic patients and enabled ranking of feature importances, as it is stable in structured data [10]. LR demonstrated the highest sensitivity, consistent with ease of interpretation in a clinical setting and a tendency to mark patients as borderline when class weights are applied, potentially resulting in a higher rate of false positives [7] [9]. XGBoost demonstrated strong discriminative ability through non-linear models and regularization, commonly resulting in stable AUC values [11]. This was achieved by the hybrid model, which aggregated probabilities, resulting in a compromise toward screening that reduced the rate of false negatives compared to tree-only methods and controlled the rate of false positives compared to the LR approach. Such ensemble behavior is consistent with established findings that combining diverse learners improves stability and reduces individual model bias [21] [22].

5.2. Practical Implications for Sierra Leone

One of the main goals of this study is its feasibility. This is because many low-resource clinics lack the ability to easily test the biomarker. For this reason, a smaller set of features, such as glucose, BMI, age, and other factors, is more feasible for early disease detection [2] [4] [8]. This is because, in such facilities, the main aim of the decision-support system is to help healthcare providers identify high-risk patients for further testing and counseling. Importantly, prioritizing recall aligns with public health needs where undiagnosed diabetes remains common and delayed detection worsens outcomes [1]-[3].

In practical deployment, the soft-voting output can be computed on low-cost hardware because it requires only running three trained models and averaging their probability outputs. The system can present clinicians with a simple risk score (hybrid probability) and a binary screening flag based on the selected threshold, alongside the key input variables. In low-resource primary healthcare units, this can support triage by identifying high-risk individuals for confirmatory testing and follow-up, while keeping computation lightweight and implementation feasible [1] [2].

6. Limitations

However, this research has a few limitations that need to be taken into account when interpreting the results. Firstly, the main data on which the models are built is a synthetic data set with a Sierra Leone perspective, designed to reflect the expected demographics and clinical parameters. Although this has helped conduct

research in a data-scarce environment while keeping privacy concerns at bay, synthetic data can sometimes overlook the clinical realities of diabetes diagnosis [1] [23].

Secondly, the set of features used is a reduced set of predictors adapted from a typical dataset used for diabetic patients. Key risk factors such as dietary habits, physical activity, social class, medication history, and comorbid conditions were excluded, which may affect the completeness of the model used for screening at the population level [2] [4].

Thirdly, although the validation of the benchmark was performed on the Pima Indians Diabetes Dataset, the chosen benchmark is a specific case and not necessarily indicative of the distribution of the population of West Africa. Thus, the validation of the benchmark serves as an indicator of robustness, but does not replace validation using real Sierra Leone clinical data or multi-site datasets [22].

Lastly, the model evaluation was conducted using only one train-test split; future work will include repeated stratified *k*-fold cross-validation to provide more robust estimates and reduce split sensitivity [24] [25].

In addition, future work will compare Random Forest feature importance with alternative interpretability approaches such as permutation importance and SHAP to strengthen explanation and transparency of model decisions in clinical use [19].

Acknowledgments

The author acknowledges Nankai University and Professor Chu for supervision, guidance, and constructive feedback throughout this research. The author also acknowledges all contributors who provided feedback and domain guidance during the research.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Magliano, D.J., Boyko, E.J., IDF Diabetes Atlas, *et al.* (2021) Covid-19 and Diabetes. In IDF Diabetes Atlas [Internet]. 10th Edition, International Diabetes Federation.

[2] World Health Organization, *et al.* (2020) World Health Statistics 2020.

[3] Atun, R., Davies, J.I., Gale, E.A.M., *et al.* (2017) Diabetes in Sub-Saharan Africa: From Clinical Care to Health Policy. *The Lancet Diabetes & Endocrinology*, **5**, 622-667.

[4] World Health Organization, *et al.* (2017) Who Country Cooperation Strategy 2017-2021: Sierra Leone.

[5] Kaviyaadharshani, D., Nivedhidha, M., Jeyarohini, R., Rani, J.L.E., Ramkumar, M.P. and Selvan, G.S.R.E. (2024) Diagnosing Diabetes Using Machine Learning-Based Predictive Models. *Procedia Computer Science*, **233**, 288-294. https://doi.org/10.1016/j.procs.2024.03.218

[6] Dutta, A., Hasan, M.K., Ahmad, M., Awal, M.A., Islam, M.A., Masud, M., *et al.* (2022) Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *Inter-*

national Journal of Environmental Research and Public Health, **19**, Article No. 12378. https://doi.org/10.3390/ijerph191912378

[7] He, H.B. and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284. https://doi.org/10.1109/tkde.2008.239

[8] Bachmann, K.N. and Wang, T.J. (2017) Biomarkers of Cardiovascular Disease: Contributions to Risk Prediction in Individuals with Diabetes. *Diabetologia*, **61**, 987-995. https://doi.org/10.1007/s00125-017-4442-9

[9] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. Wiley. https://doi.org/10.1002/9781118548387

[10] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[11] Chen, T.Q. (2016) Xgboost: A Scalable Tree Boosting System. Cornell University.

[12] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[13] UCI Machine Learning (2016) Pima Indians Diabetes Database. https://kaggle.com/uciml/pima-indians-diabetes-database

[14] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C. and Johannes, R.S. (1988) Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Washington DC, 6-9 November 1988, 261.

[15] Dua, D. and Graff, C. (2019) UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 1-29. http://archive.ics.uci.edu/ml

[16] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., *et al.* (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[18] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[19] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *NIPS*'17: *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[20] Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G.M. (2015) Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *British Journal of Surgery*, **102**, 148-158. https://doi.org/10.1002/bjs.9736

[21] Dietterich, T.G. (2000) Ensemble Methods in Machine Learning. In: Kittler, J. and Roli, F., Eds., *International Workshop on Multiple Classifier Systems*, Springer, 1-15. https://doi.org/10.1007/3-540-45014-9_1

[22] Kuncheva, L.I. (2014) Combining Pattern Classifiers: Methods and Algorithms. Wiley. https://doi.org/10.1002/9781118914564

[23] Alberti, K.G.M.M. and Zimmet, P.Z. (1998) Definition, Diagnosis and Classification of Diabetes Mellitus and Its Complications. Part 1: Diagnosis and Classification of

Diabetes Mellitus. Provisional Report of a WHO Consultation. *Diabetic Medicine*, **15**, 539-553.
https://doi.org/10.1002/(sici)1096-9136(199807)15:7<539::aid-dia668>3.0.co;2-s

[24] Omar, R. (2010) Clinical Prediction Models: A Practical Approach to Development, Validation and Updating by Steyerberg, E.W. *Biometrics*, **66**, 661-662.
https://doi.org/10.1111/j.1541-0420.2010.01431.x

[25] Kohavi, R., *et al.* (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, Volume 14, 1137-1145.