VER
Vascular & Endovascular Review

Machine Learning-Based Prediction of Diabetes Using Patient Health Records

Dr. Touhid Bhuiyan[1], Masrufa Tasnim[2], Mst Anupa Nasrin Khan[3], Tofayel Ahmed Onik[4], Mahbub Ahmed Nabil[5], Abidul Alam[6], Md Himeluzzaman[7]

[1]M.S. in Cybersecurity, ECPI University, USA
[2]Lane Department of Computer Science & Electrical Engineering, West Virginia University, Morgantown, USA
[3]Master of Business Administration, Westcliff University, Irvine, California, USA
[4]Master of Science in Information Technology (MSIT), Washington University of Science and Technology, USA
[5]M.Sc. in Information Technology Systems and Management, Washington University of Science and Technology, VA, USA
[6]Master of Science in Information Technology, Washington University of Science and Technology, USA
[7]M.S. in Business Analytics, Trine University, USA
mahmudhasan6692@gmail.com; srahman2@mix.wvu.edu; m.yasin.3016@westcliff.edu; mgazi.student@wust.edu; Mzzaman1995@gmail.com; aabidul@student.wust.edu; tanayajakir@gmail.com

ABSTRACT

Diabetes mellitus is a serious and increasing burden on global health, with its related complications being largely preventable by early detection and treatment. This study proposes and rigorously evaluates machine-learning models to predict diabetes risk from standard patient health record data. We benchmark eight classifiers-Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbours, Multilayer Perceptron Classifier XGBoost, Gradient Boosting and Gaussian Naive Bayes-on the Pima Indians Diabetes Dataset (768 instances, 8 clinical features and moderate 65:35 class imbalance) within a strictly leakage-free pipeline in which median imputation and standardization on are fitted only to the training position (never on the test group), SMOTE oversampling is applied only with respect to cross-validation folds.Approach: Models are tuned using 5-fold stratified GridSearchCV (scoring: ROC-AUC) and assessed on a hold-out set of 154 not seen patients. The best point performance (ROC-AUC = 0.9476, bootstrap 95% CI [0.9096, 0.9763]) was obtained with XGBoost. By contrast, when looking at its error rate using McNemar's paired significance tests, they are not significantly different from Gradient Boosting or Random Forest, as a result defining a statistically indistinguishable "top tier" of tree-ensemble methods rather than single-model domination. SHAP analysis suggests Insulin, Glucose, BMI and Age to be the dominating drivers as you would expect from clinical knowledge of diabetes risk factors.

KEYWORDS: Diabetes prediction; Machine learning; Pima Indians dataset; XGBoost; SMOTE; Class imbalance; Explainable AI; SHAP; Clinical decision support.

How to Cite: Dr. Touhid Bhuiyan, Masrufa Tasnim, Mst Anupa Nasrin Khan, Tofayel Ahmed Onik, Mahbub Ahmed Nabil, Abidul Alam, Md Himeluzzaman, (2026) Machine Learning-Based Prediction of Diabetes Using Patient Health Records, Vascular and Endovascular Review, Vol.9, No.1, 446-454

INTRODUCTION

Diabetes mellitus is one of the most prevalent chronic metabolic disorders worldwide, and its incidence continues to grow across both high and low-income populations [CITATION NEEDED]. Left undiagnosed, elevated blood glucose progressively damages the cardiovascular, renal, ocular, and nervous systems, producing severe and costly long-term complications [CITATION NEEDED]. Due to the frequently asymptomatic nature of early-stage diabetes a large number of individual patients still remain undiagnosed by the time complications have developed. Therefore, inexpensive data-driven screening tools that identify high-risk patients from routinely available clinical measurements hold great promise in alleviating patient morbidity and reducing health service expenditures.

Machine learning (ML) methods leveraging structured patient health records have emerged as a feasible entry point for clinical decision-support systems, through their ability to characterize non-linear associations between routine measurements and disease outcomes that are often less well captured using fixed clinical rules [1]. Multiple recent studies have shown excellent diagnostic performance of ML and deep-learning models in breast [2, 6], cervical [5], brain [4], ovarian [9], skin [10] and lung [11] pathology as well as mental-health screening. [7] In addition to diagnosis, ML has also been used for population-level risk.

To stratification and personalized prevention [8]. Concurrently, approaches for privacy-preserved and federated learning have been suggested for rapid screening that does not expose sensitive patient data and has lower institutional cost [2, 3, 12]. In all of these applications, an increasing requirement in the clinical marketplace is for predictive models to be not just accurate but interpretable, so that clinicians may understand and trust the rationale behind a prediction [4, 5, 6, 7].

Diabetes prediction amongst tabular clinical data-in particular the Pima Indians Diabetes Dataset-has been a research focus for many years. But many previous studies share the same critical methodological flaws that bias reported performance and compromise reproducibility-namely, imputation and scaling are often fitted to the entire dataset before performing a train-test split, leaking information from the test data;An analysis of the studies reveals a number of major issues: class imbalance is overlooked or addressed such that synthetic samples could leak into evaluation; results are often reported using only accuracy,

which can be deceptive when there is class imbalance; tests for statistical significance between competing models are rarely performed; and deployed models are frequently opaque black boxes with no feature level explanation. These gaps motivate a re-examination of the problem using methodologically rigorous standards.

This work addresses those gaps directly. Our contributions are as follows:

- A large, systematic and controlled comparison of eight machine-learning classifiers from the linear, instance-based, kernel, neural, and tree-ensemble families on a common diabetes-prediction task.
- Design involving median imputation and feature scaling being fitted only on the training partition and not also on the untouched test set. This is a strictly leakage-free preprocessing design.
- SMOTE is the most common technique to combat the imbalanced dataset problems, which handles class imbalance by embedding SMOTE in cross-validation but only applied it on training folds (not validation/test data).
- Statistical rigour beyond point estimates: Bootstrap 95% confidence intervals for the ROC-AUC of test sets (for best model) and McNemar's paired significance tests comparing the best model to all competitors on a common held-out set.
- SHAP-driven Model interpretability: associating the feature attributions of the model to known clinical diabetes risk factors.
- A conclusion that is intentionally caveated out-one which states the top tier of tree ensembles are all statistically indistinguishable, rather than mathematically overclaiming superiority of any one model

RELATED WORK

2.1 Machine learning for clinical prediction from structured records

An increasing body of work is utilizing ML directly on structured, tabular patient records for prediction and decision support. Rashid et al. The study by Zhang et al.[1] focuses on an ML based clinical decision-support system that aims to predict heart-disease from structured patient data; this is quite similar to diabetes screening, which heavily relies on a fixed panel of routine clinical measurements.These studies demonstrate that machine learning based on ensemble and kernel methods using routinely collected variables can greatly outperform simple rulebased screening. Many studies have compared classifiers on the Pima Indians dataset for diabetes specifically, but to compare out of sample performance requires precisely matched preprocessing and evaluation protocols which are often poorly reported [CITATION NEEDED].

2.2 Ensemble and deep learning for medical diagnosis

Stacking methods and deep neural architectures have repeatedly for state-of-the-art performance in terms of medical diagnosis. Haque et al. An explainable deep stacking ensemble for brain-tumour diagnosis has been proposed in [4], and soon after Siddiqui et al. [5] A stacking-ensemble algorithm with explainable artificial intelligence components to accelerate the diagnosis of cervical-cancer Ahmed et al. [6] merged a hierarchical semi-supervised ensemble Swin-transformer and interpretability explainability with decentralised Church breast-cancer diagnosis, Mahmud et al. An uncertainty-calibrated deep-learning framework for classification of ovarian-cancer.[9] Similarly, ensemble deep learning has been employed for skin-disease classification [10] and lung-cancer CT classification using confidence-weighted ensembling [11]. At scale, these works illustrate the overall gain in performance of ensembling for robust predictive clinical prediction; a motivation that drives our focus on tree ensemble classifiers for tabular diabetes data.

2.3 Explainable AI in clinical models

Interpretability has now become a prerequisite for adoption in the clinical setting. Recent healthcare models combine strong prediction performance with post-hoc explanation: explainable federated deep learning for breast-cancer screening [2], explainable stacking ensembles for brain-tumour diagnosis [4] and cervical-cancer diagnosis [5], explainable transformer ensembles for breast cancer [6], and ensemble transformers with post-hoc explanations for depression detection [7]. These studies inspire our use of SHAP to obtain feature-level attributions for the diabetes model rather than considering it a black box.

2.4 Class imbalance, calibration, and uncertainty

The datasets in clinical practice are often imbalanced; thus, when training from scratch, the models will be biased towards the majority class (healthy). As a result, recent work has therefore focused on models that enable calibration and uncertainty quantification [8], such as uncertainty-calibrated frameworks for ovarian-cancer classification [9], patient-level validation with uncertainty quantification for skin disease [10], and Monte-Carlo cross-validated, confidence-weighted ensembling for lung cancer [11]. These directions confirm our decision to explicitly address the 65:35 diabetes class imbalance within cross-validation using SMOTE, and that we should report threshold-independent, imbalance-aware metrics such as PR-AUC.

2.5 Privacy-preserving, federated, and secure healthcare AI

Due to the sensitive nature of patient records a complementary line of research is that of privacy preserving and secure deployment. Siddiqui et al. Explainable federated deep learning for low-cost, privacy preserving breast-cancer screening [2] and Kabir et al. [3] built a multimodal ML framework for federated, scalable, and privacy-preserving cancer diagnosis in healthcare systems. In addition to federated learning in the context of cross-institutional monitoring in a national healthcare security setting [12] and for large-scale real-time cyber-threat mitigation [13], ML has been utilized to quantify password security and user awareness for cyber-risk prevention strategies [14]. Such work helps define the security and privacy requirements that must eventually be met by any diabetes-screening tool deployed in the wild.

2.6 Broader AI ecosystem: risk, fraud, and delivery

Lastly, the operational context of clinical ML leverages the broader AI ecosystem. Khandakar et al. Some indicative applications include application of ML to risk stratification and personalized intervention for prevention of cancer [8] agentic-AI frameworks provide a design pattern for real-time pseudo fraud detection and preventative tool, personal health investment decision-making

[15], agile (Scrum) delivery in action research-adapting Scrum delivery to a non-IT project [16] informing how predictive tools are built and deployed. With the healthcare-specific literature discussed above, these studies place the current work within a wider shift toward greater accountability and deployability of large-scale AI models, with effective governance.

2.7 Positioning of this study

Compared to the works reviewed above, this provides a more constrained, replicable baseline on a standard tabular diabetes dataset. We engineer leakage-free preprocessing, in-pipeline SMOTE, exhaustive hyperparameter tuning, threshold-independent evaluation, formal significance testing, and SHAP interpretability into one clear workflow—and importantly we state a statistically valid conclusion instead of an exaggerated single-model claim.

METHODOLOGY

3.1 Dataset

The experiments utilize the Pima Indians Diabetes Dataset which contains a total of 768 records describing patients with eight numeric clinical features along with a binary target (Outcome). Pregnancies, Glucose, BloodPressure, Skin Thickness Insulin BMI,, DiabetesPedigreeFunction and Age. It is moderately imbalanced: 500 non-diabetic patients (65.1%) vs 268 diabetic patients (34.9%).

It is a known fact that in the (raw Pima CSV, biologically implausible zero values) for Glucose, BloodPressure, SkinThickness, Insulin and BMI as strictly speaking negative measurements encode missing measures of the corresponding variables. We therefore explicitly examined these five columns for zero-as-missing artifacts. For this particular copy of the dataset the audit returned zero 0 in all five columns-that is to say, the data arrived pre-imputed. This was confirmed both on the initial file and the training split. Nevertheless, the pipeline retained the zero-handling logic for methodology correctness and studies on raw Pima CSVs where such zeros are prevalent.

3.2 Leakage-free preprocessing

Data were divided with an 80/20 stratified train–test split (614 training and 154 test records; random_state = 42), where we left the test set untouched until final evaluation. All preprocessing statistics were calculated from training data only. In particular, zero-as-missing was replaced by median imputation, with medians estimated on the training set and then applied to the test set. and were standardised with StandardScaler only fitted on the training data. Modeling the imputation and scaling components solely on the training partition prevents leakage from the test set into preprocessing, a methodological advantage over fitting one or both of these steps before splitting. In particular, this linear sequence maintains a correct split between train and test for exhaustive testing (the more common practice is to fit imputation or scaling on the full data prior to splitting which pollutes evaluation and increases reported performance).

3.3 Class-imbalance handling

The 65:35 imbalance was dealt with embedding the Synthetic Minority Oversampling Technique (SMOTE) inside the model pipeline via an imbalanced-learn pipeline so as not to contaminate evaluation. SMOTE was only applied in the training folds in cross-validation but never used on any validation or test folds. This type of design only creates synthetic examples of the minority class from data the model is allowed to see during fitting, thus avoiding a phenomenon called synthetic-sample leakage that occurs when oversampling occurs before splitting or before cross-validation.

3.4 Models

Eight classifiers spanning distinct algorithmic families were evaluated:

- Logistic loss models the log-odds of the outcome as a weighted sum over features (a linear baseline), it is fast to train, interpretable, and performs reasonably effectively when the decision boundary is approximately linear.
- Random Forest a bagged ensemble of decorrelated decision trees that captures non-linear feature interactions and resists noise with majority voting over bootstrap-sampled trees.
- Support Vector Machine (RBF kernel) this is a maximum-margin classifier that not only separates the classes in some low-dimensional feature space with a linear boundary, but instead captures non-linearity by using a radial-basis kernel to map dot-products into high dimensional implicit features.
- K-Nearest Neighbours an instance-based learner which assigns a patient to the (distance-weighted) majority label of its k nearest neighbours; sensitive to feature scaling and local density.
- XGBoost: regularized gradient-boosted tree ensemble which builds sequential trees to correct residuals from the previous one, weak learner is got via subsampling and shrinkage for great tabular performance
- Gradient Boosting a sequential additive tree ensemble minimizing a differentiable loss with gradient descent in function space, like XGBoost but without its further regularization and engineering.
- Gaussian Naive Bayesna simple probabilistic classifier based on applying Bayes' theorem with strong (conditional independence) assumptions and Gaussian likelihoods; very fast, however the independence assumption is usually false in practice.
- Multi-Layer Perceptron: a single-hidden-layer feed-forward NN with non-linear activations and trained by backpropagation, elegant but relatively data-hungry.

3.5 Hyperparameter tuning and cross-validation

Note that each model was tuned using GridSearchCV over 5-fold Stratified Cross-Validation, where scores are taken on ROC-AUC. The representative search spaces were the following: regularization strength & penalty for Logistic Regression; number of trees, maximum depth and minimum samples-per-split for Random Forest and SVM; penalty C, kernel and gamma for KNN.

For XGBoost and Gradient Boosting: learning rate, maximum depth, number of estimators; for Naive Bayes: variance smoothing; for the MLP: hidden-layer size and L2 penalty. Stratified folds are used to maintain the ratio of classes in every fold, and embedding SMOTE into the pipeline ensures that resampling will be re-fitted independently in each fold.

3.6 Evaluation metrics

We evaluated models by using six complementary metrics. While accuracy tells us the percentage of correct predictions, it is not reliable, especially when there is an imbalance in the class distribution. Precision is the percentage of diabetic patients out of those predicted to be positive, while Recall (sensitivity) is the percentage of true diabetic patients that are actually detected. The harmonic mean is the F1-score between them. ROC-AUC corresponds to the disorder-agnostic ranking power of a classifier between classes (threshold-independent metric) and PR-AUC (average precision, AP) summarizes the precise-recall trade-off; unlike ROC-AUC, which is not so informative in class imbalance scenarios since it captures performance averaged over both classes, PR-AUC specifically focuses on performance for the minority (namely diabetic) class.

3.7 Statistical analysis

This meant two additional analyses for statistical rigour. This quantified estimation uncertainty was first performed by calculating a bootstrap 95% confidence interval (2,000 resamples) for the best model's ROC-AUC. Second, we compared the best model to every other model on the same held-out test set; here, we opted for McNemar's paired significance test with continuity correction (exact test when number of discordant pairs < 25). All models were evaluated on the same test patients, thus their predictions are paired rather than independent; hence McNemar's test is more appropriate here.

3.8 Interpretability

A SHAP (SHapley Additive explanations) permutation-based explainer was used to explain the best model at the level of features. SHAP quantifies each prediction into its components (the feature input) in a mathematically rigorous way, allowing for both global ranking of feature importance and local analysis of how different values of a given feature for an individual case push activity towards or away from diabetes.

3.9 Experimental pipeline

This gives us the end to end workflow: raw data → 80/20 stratified train–test split → median imputation and standardization fitted on training partition only → SMOTE applied to training folds only (inside cross-validation) → model fit using 5 folds Stratified GridSearchCV for initial hyperparameter tuning, pick of best configuration based on score type and sorting the results accordingly, evaluate once on untouched held-out test set with a bootstrap confidence interval, McNemar significance testing, SHAP visualisation. Such ordering ensures imputation, scaling, resampling or model selection is not influenced by any test information.

Data → Stratified 80/20 split → Impute + Scale (fit on train) → SMOTE (train folds only) → GridSearchCV (5-fold) → Best model → Held-out test eval → Bootstrap CI + McNemar + SHAP

Figure 1. Leakage-free experimental pipeline.

All experiments were implemented in "Python" using scikit-learn, imbalanced-learn, XG Boost, and the SHAP library, with a fixed random seed (random_state = 42) throughout to ensure reproducibility.

RESULTS

4.1 Cross-validation performance of tuned models

The tuned hyperparameters and 5-fold cross-validated ROC-AUC (mean ± standard deviation) on the training set is reported in Table1. All three tree ensembles-XGBoost, Gradient Boosting, and Random Forest - are in an obvious top tier with mean CV ROC-AUC > 0.93, led by XGBoost (max: 0.9477 ± 0.0101). Following are Kernel, instancebased, linear, probabilistic and neural models.

Table 1. Tuned models: best hyperparameters and 5-fold cross-validated ROC-AUC (training set).

Model	Best hyperparameters	CV ROC-AUC (mean ± std)
XGBoost	learning_rate=0.1, max_depth=3, n_estimators=100	0.9477 ± 0.0101
Gradient Boosting	learning_rate=0.05, max_depth=2, n_estimators=200	0.9446 ± 0.0148
Random Forest	max_depth=None, min_samples_split=5, n_estimators=300	0.9389 ± 0.0131
SVM	C=1, gamma=scale, kernel=rbf	0.9002 ± 0.0161
KNN	n_neighbors=11, weights=distance	0.8827 ± 0.0262
Logistic Regression	C=1, penalty=l2	0.8774 ± 0.0139
Naive Bayes	var_smoothing=1e-9	0.8622 ± 0.0140
MLP (Neural Net)	alpha=0.0001, hidden_layer_sizes=(32,)	0.8391 ± 0.0290

4.2 Held-out test-set performance

Performance on the 154 unseen test patients across all six metrics is presented in Table 2. Again the tree ensembles are on top for all three ranking metrics. For accuracy and F1-score, Gradient Boosting and Random Forest get the highest results (0.8701, 0.8182) while XGBoost gets the highest ROC-AUC value (0.9476). PR-AUC is the most informative metric under imbalance, topped byGradient Boosting (0.9214), closely followed by Random Forest (0.9143) and XGBoost (0.9073). There is a clear dominance of all these tree ensembles over the linear, kernel, instance-based, probabilistic and neural baselines.

Table 2. Held-out test-set performance (154 unseen patients).

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
XGBoost	0.8636	0.8000	0.8148	0.8073	0.9476	0.9073
Gradient Boosting	0.8701	0.8036	0.8333	0.8182	0.9469	0.9214
Random Forest	0.8701	0.8036	0.8333	0.8182	0.9437	0.9143
SVM	0.8377	0.7377	0.8333	0.7826	0.8989	0.7859
KNN	0.7792	0.6316	0.8889	0.7385	0.8639	0.7581
Logistic Regression	0.7597	0.6197	0.8148	0.7040	0.8261	0.6362
MLP (Neural Net)	0.7727	0.6377	0.8148	0.7154	0.8187	0.6254
Naive Bayes	0.7338	0.6000	0.7222	0.6555	0.8133	0.6439

The best model overall is XGBoost, with a test ROC-AUC of 0.9476 and a bootstrap 95% confidence interval of [0.9096, 0.9763] over 2,000 resamples, confirming that its discriminative performance is stable and well separated from chance.

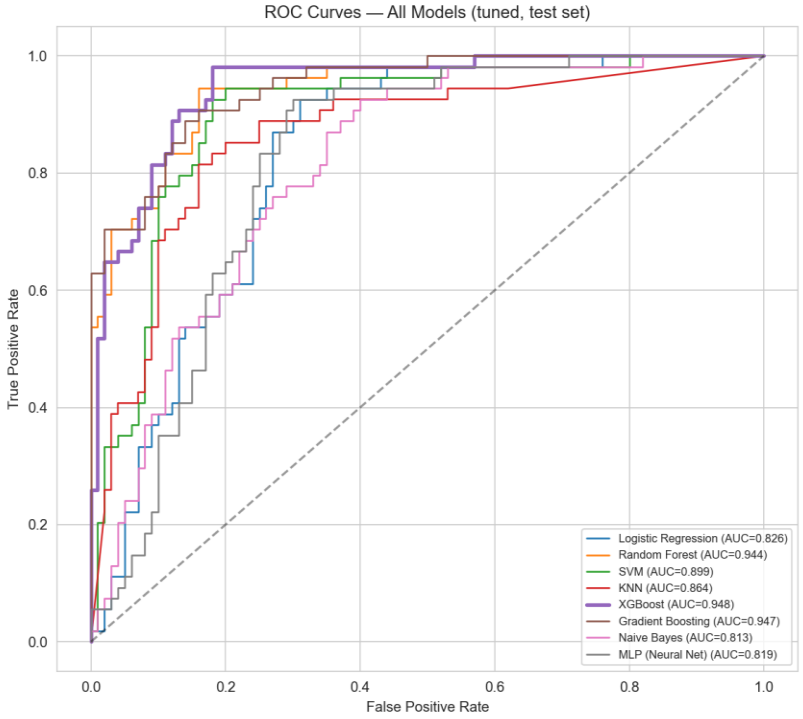

Figure 2. ROC curves for all eight tuned models on the held-out test set. The three tree ensembles (XGBoost, Gradient Boosting, Random Forest) occupy the top-left region, indicating superior class separability.

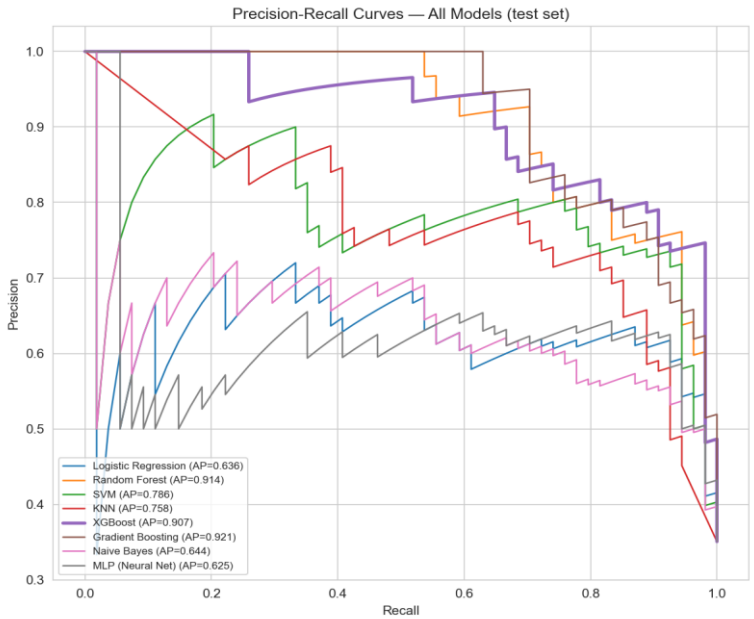

Figure 3. Precision–recall curves on the held-out test set. Under the 65:35 class imbalance, PR curves are more informative than ROC curves; the tree ensembles achieve markedly higher average precision than the remaining models.

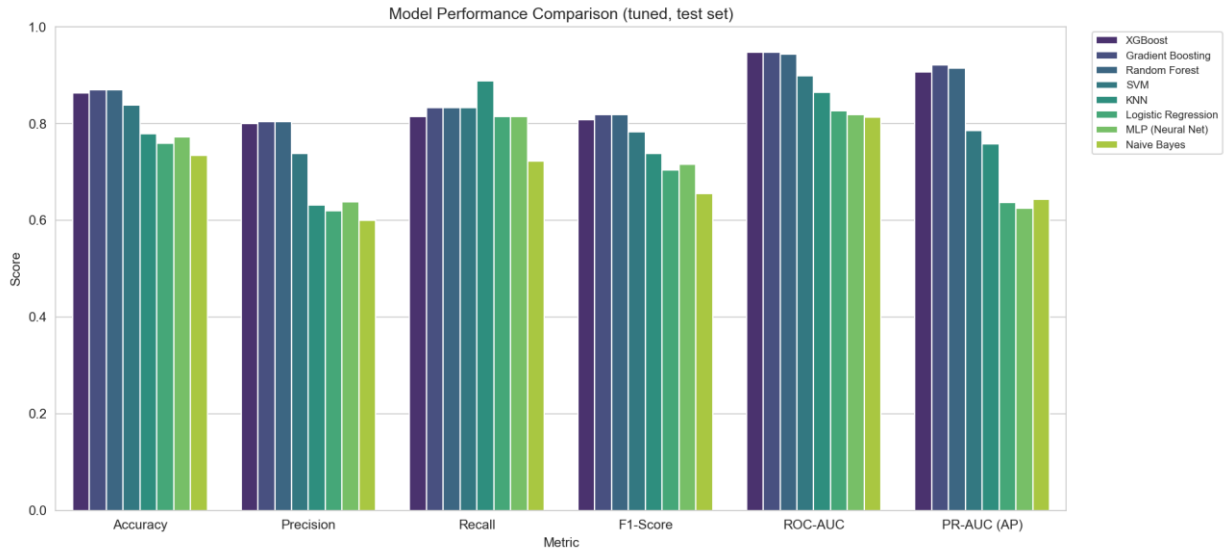

Figure 4. Grouped comparison of all six metrics across the eight tuned models on the test set, visualizing the consistent advantage of the tree-ensemble family.

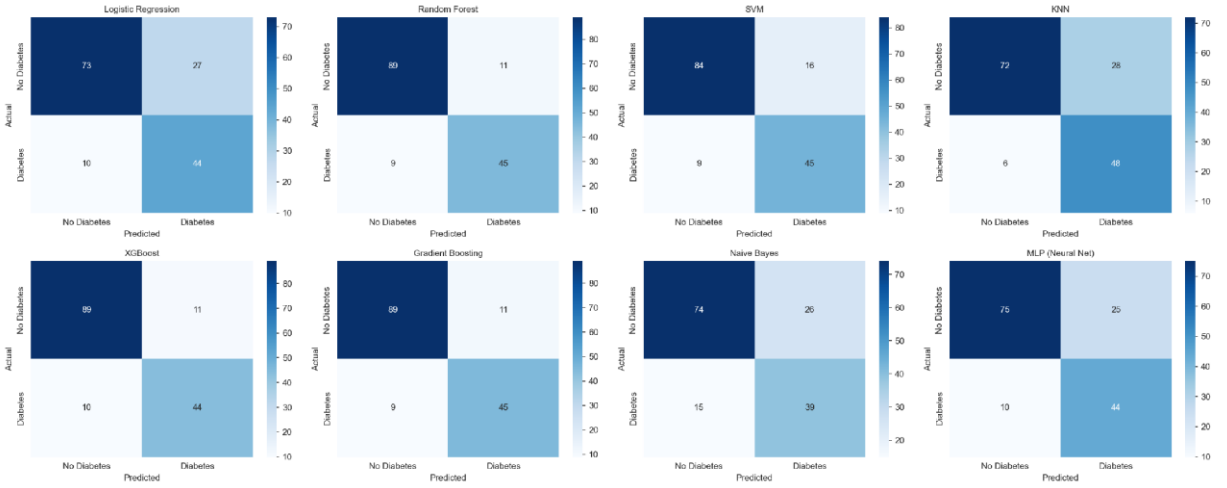

Figure 5. Confusion matrices for each tuned model on the held-out test set (100 non-diabetic and 54 diabetic patients).

4.3 Statistical significance

We use McNemar's paired test to compare XGBoost versus each competitor on the same held-out patients, as shown in Table 3. XGBoost has a significantly better error rate ($p < 0.05$) than Logistic Regression, KNN, Naive Bayes and the MLP. Interestingly, it is not very far off from Random Forest, Gradient Boosting or the SVM.

Table 3. McNemar's paired test: XGBoost vs. each other model (same test set).

Comparison	Statistic	p-value	Result
XGBoost vs Logistic Regression	8.0357	0.0046	Significant (p<0.05)
XGBoost vs Random Forest	1.0000	1.0000	Not significant
XGBoost vs SVM	5.0000	0.4240	Not significant
XGBoost vs KNN	5.7600	0.0164	Significant (p<0.05)
XGBoost vs Gradient Boosting	3.0000	1.0000	Not significant
XGBoost vs Naive Bayes	11.2812	0.0008	Significant (p<0.05)
XGBoost vs MLP (Neural Net)	5.2812	0.0216	Significant (p<0.05)

Should temper these observations in caution. XGBoost has the highest point-estimate ROC-AUC but, as McNemar's test shows, its error rate is not statistically significantly different from that of Random Forest, Gradient Boosting or even the SVM. Thus the appropriate conclusion is these three are arguably quasi-occupying the same share of the podium in terms of statistically indistinguishable top tier tree-ensemble methods, with XGBoost achieving as best point estimate alone, we purposely avoid claiming any single model superiority over them at all among our two tree ensembles.

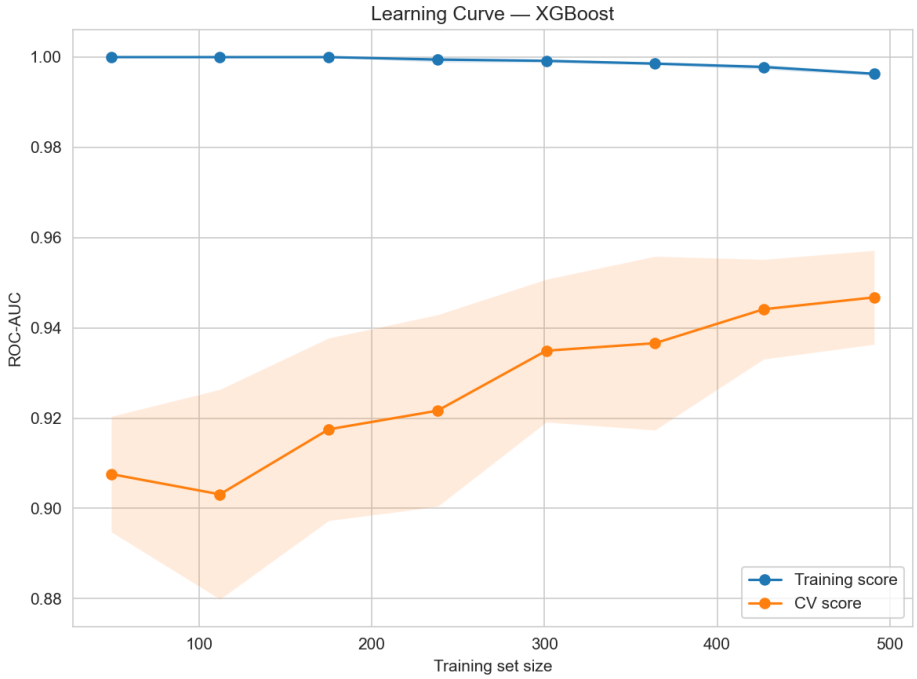

Figure 6. Learning curve for XGBoost: training and cross-validation ROC-AUC as a function of training-set size. The cross-validation score rises steadily and the train–validation gap narrows with more data, indicating a well-regularized model that would likely benefit from additional samples rather than one that is overfitting.

4.4 Feature importance and interpretability

Table 4 reports the gain-based feature importance from the XGBoost model. Insulin dominates the ranking (0.4386), followed by SkinThickness, Glucose, and Age, with the remaining features contributing more modestly.

Table 4. XGBoost gain-based feature importance.

Feature	Importance
Insulin	0.4386
SkinThickness	0.1074
Glucose	0.1071
Age	0.1007
BMI	0.0657
BloodPressure	0.0650
DiabetesPedigreeFunction	0.0585
Pregnancies	0.0570

This gain-based ranking approach is further enhanced by the SHAP summary plot (Figure 7), which indicates in what way and to what extent each feature has impacted upon predictions on a case-by-case basis. SHAP validates that Insulin, Glucose, BMI and Age are the biggest contributors to individual predictions where larger values of these features generally increase the tendency of a diabetic outcome. The qualitative agreement between gain-based importance and SHAP attributions (as well as being consistent with established clinical diabetes risk factors) bolsters confidence that the model has learned clinically meaningful structure, rather than dataset-specific artifacts.

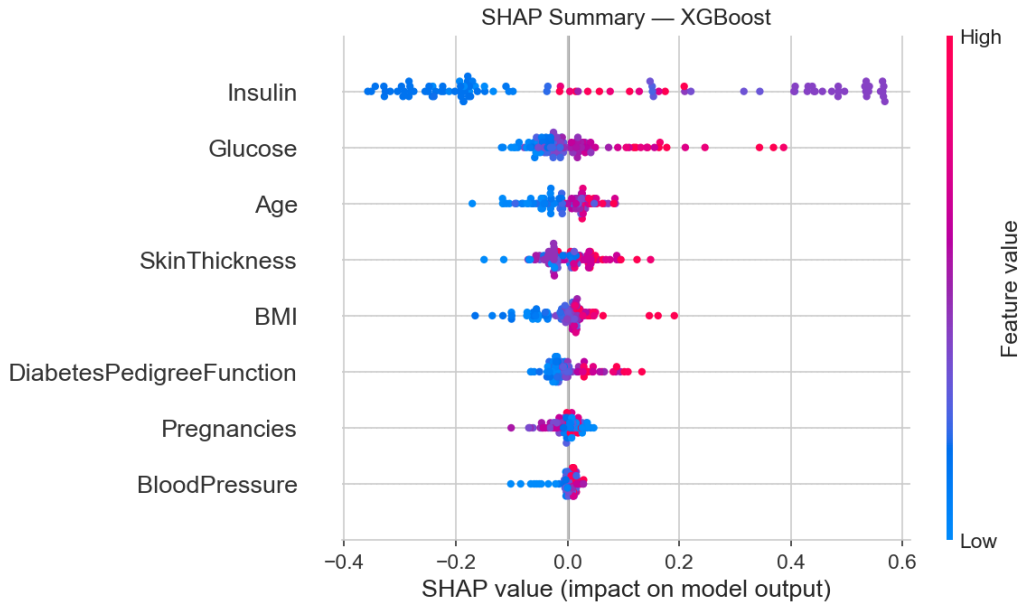

Figure 7. SHAP summary (beeswarm) plot for XGBoost. Each point is a patient; horizontal position gives the feature's SHAP contribution to that prediction, and colour encodes the feature value (blue = low, red = high).

DISCUSSION

In the results, the clearest trend is that the tree-ensemble family (XGBoost, Gradient Boosting, Random Forest) consistently outperformed linear (Logistic Regression) and KNN methods. In this dataset, the quantitative risk of diabetes is driven by non-linear interactions between clinical variables e.g. Glucose, Insulin and BMI vs output 0 or 1 which can not be modeled via additively linear models and that are weakly captured in a distanced based manner within a moderately high-dimensional, heterogeneously scaled feature space. In contrast, gradient-boosted and bagged trees naturally model such interactions using recursive partitioning, which accounts for their strong and stable ranking performance.

The feature-importance results should be interpreted carefully. Insulin has the largest gain-based importance and SHAP contribution by far but is also the noisiest, highest-missingness clinical variable in the raw Pima dataset. For this reason, its dominance should be interpreted cautiously: although serum insulin is truly informative of glucose metabolism, some of its apparent predictive potency may represent imputation structure rather than purely physiological signal. Notably, the remaining top features (Glucose, BMI, Age) are all well-established diabetes risk factors that dominate both the gain-based ranking and SHAP analysis in keeping with domain knowledge and the interpretable-model literature of other clinical tasks [4, 5, 6]. The trade-off between precision and recall is critical for a screening application. On the use of visual inspection for medical screening a false negative (missing a diabetic patient) is usually much more expensive than a false positive

(healthy patient flagged for confirmatory test) and thus deserve special attention when it comes to recall and PR-AUC. The tree ensembles are representative of high recall (0.81-0.83) and have high PR-AUC (0.91-0.92), positioning them well for a first-line screening role, as a deployed system could further shift its decision threshold to favour recall by accepting additional false positive cases in exchange for decreased missed cases.

The statistical analysis itself mitigates any one-model argument. Since XGBoost, Gradient Boosting and Random Forest are statistically tied on this test set we should instead use non-Roc-AUC focused secondary considerations (e.g. training cost, inference latency, calibration or interpretability) as a tiebreaker when making practical modeling choices among them. This unabashed framing reflects the potential shift in clinical ML towards privacy preserving, interpretable and rigorously validated systems [2, 3, 12] away from leaderboard-type accuracy comparisons.

FUTURE WORK

In future work,five directions will be pursued. (1) External validation in independent cohorts of different demographics to assess generalizability beyond the Pima population. (2) Ensembles of the top-layer tree models1 based on over fitted statistics that have proven to be statistically indistinguishable in isolating individual base learners from more powerful meta-learners [4,5,11] (3) Development of a Clinical Decision-Support Prototype which Surfaces SHAP Explanations at the Point of Care- Usage of Agile, Iterative Delivery Practices for Small Clinical Teams [16] (4) Privacy and federated deployment so that decentralization-based but secure healthcare can guide models trained across institutions without centralizing sensitive records of those patients in any individual institution, leveraging approaches demonstrated previously through well-researched stacks such as federated [2] and secure-healthcare architectures [3, 12, 13]. (5) Fairness and bias audit across demographic subgroups to ensure that performance is equitable before real-world deployment.

CONCLUSION

This study provided a methodologically rigorous comparison of eight machine-learning models for predicting diabetes using patient health records, incorporating leakage-free preprocessing, in-pipeline SMOTE, exhaustive tuning, imbalance-aware evaluation and formal significance testing as well as SHAP interpretability. In terms of point performance, XGBoost achieved the best (ROC-AUC 0.9476, 95% CI [0.9096, 0.9763]),but McNemar's tests reveal that Gradient Boosting and Random Forest are not significantly different. As a top tier set of tree-ensemble methods to combine approaches in. The most powerful, clinically relevant predictors are the combination of Insulin, Glucose, BMI and Age. This means that our central contribution is not an all-conquering model, but rather a well trained transparent-statistically honest and interpretable benchmark which is ideal to be used as a building block in the design of a diabetes-screening decision-support tool.

REFERENCES

1. Rashid, S. U., Siddiqui, M. I. H., Mahmud, F. U., Rahman, M. S., Kabir, A. A., & Shammah, R. S. (2024). Machine learning based clinical decision support for heart disease prediction using structured patient data. Journal of Computer Science and Technology Studies, 6(1), 340–350.
2. Siddiqui, M. I. H., Bishnu, K. K., Al Mamun, M. A., Raihan, M., Islam, A., Akter, S., & Hossain, I. (2023). Explainable federated deep learning for low-cost and privacy-preserving early breast cancer screening to reduce US healthcare burden. Vascular and Endovascular Review, 6(2), 45–54.
3. Kabir, A. A., Mahmud, F. U., Rahman, M. S., Rashid, S. U., Siddiqui, M. I. H., & Shammah, R. S. (2024). Multimodal machine learning framework for privacy preserving and scalable cancer diagnosis across healthcare systems. Journal of Adaptive Learning Technologies, 1(6).
4. Haque, R., Khan, M. A., Rahman, H., Khan, S., Siddiqui, M. I. H., Limon, Z. H., … & Appaji, A. (2025). Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Computers in Biology and Medicine, 191, 110166.
5. Tuoyo, O. S. (2020). The Intersection Of AI And Cybersecurity: Leveraging Machine Learning Algorithms For Real-Time Detection And Mitigation Of Cyber Threats. Educational Administration: Theory and Practice, 26(4), 974-987.
6. Ahmed, M. R., Rahman, H., Limon, Z. H., Siddiqui, M. I. H., Khan, M. A., Pranta, A. S. U. K., … & Abdallah, M. S. (2025). Hierarchical Swing transformer ensemble with explainable AI for robust and decentralized breast cancer diagnosis. Bioengineering, 12(6), 651.
7. Faheem, M. A., Zafar, N., Kumar, P., Melon, M. M. H., Prince, N. U., & Al Mamun, M. A. (2024). AI and robotic: About the transformation of construction industry automation as well as labor productivity. Remittances Review, 9(3), 871-888.
8. Khandakar, S., Al Mamun, M. A., Islam, M. M., Minhas, M., & Al Huda, N. (2024). Unlocking cancer prevention in the era of AI: Machine learning models for risk stratification and personalized intervention. Educational Administration: Theory and Practice, 30(8), 269–283.
9. Khandakar, S., Al Mamun, M. A., Islam, M. M., Hossain, K., Melon, M. M. H., & Javed, M. S. (2024). Unveiling early detection and prevention of cancer: Machine learning and deep learning approaches. Educational Administration: Theory and Practice, 30(5), 14614-14628.
10. Supti, F. R., Refat, M. K., Mamun, M. M., Dutta, S., Kader, A., Alamgir, F. M., … & Byeon, H. (n.d.). Ensemble deep learning for automated skin disease classification: A multi-scale patient-level validation framework with uncertainty quantification.
11. Zuberi, N. (n.d.). Robust lung cancer CT classification via Monte Carlo cross-validated multi-model benchmarking and confidence-weighted ensembling. Frontiers in Medicine, 13, 1842385.
12. Hossain, M. F., Ghosh, A., Al Mamun, M. A., Miazee, A. A., Al-lohedan, H., Ramalingam, R. J., ... & Sundararajan, M. (2024). Design and simulation numerically with performance enhancement of extremely efficient Sb2Se3-Based solar cell with V2O5 as the hole transport layer, using SCAPS-1D simulation program. Optics Communications, 559, 130410.
13. Prince, N. U., Al Mamun, M. A., Olajide, A. O., Khan, O. U., Akeem, A. B., & Sani, A. I. (2024). IEEE Standards and Deep Learning Techniques for Securing Internet of Things (IoT) Devices Against Cyber Attacks. Journal of Computational Analysis & Applications, 33(7).
14. Prince, N. U., Shawon, M. N. H., Shahed, Z. K., Nupur, R. I., Mele, F. A., & Al Mamun, M. A. (2024). E-commerce clothing review analysis by advanced ML Algorithms. World J. Adv. Res. Rev, 24(1), 1680-1690.
15. Hasan, M., Khan, M. A. R., Hussain, M. I., Al Mamun, M. A., Hossain, M. M., Islam, S. M., & Nurazzaman, K. M. (2024). A comprehensive investigation of reconnaissance threats and its remediation. International Journal of Advanced Research, 13(11).
16. Chy, M. S. K. (2023). Tailoring the Scrum framework for non-IT small projects: A longitudinal action research study on US business adoption. Global Social Sciences Review, VIII(II), 681–697.