International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7, 2025
https://theaspd.com/index.php

Diabetes Prediction on Pima Indians Dataset Using Machine Learning Techniques

ABDELMGEID A. ALI[1], GALAL R. GALAL[2] AND HASSAN S. HASSAN[3]

[1]Faculty of Computers and Information, Minia University, Minia 61519, Egypt

[2]Faculty of Computers and Information, Minia University, Minia 61519, Egypt

[3]Faculty of Computers and Information, Minia University, Minia 61519, Egypt

***Abstract**: Type 2 diabetes is a major public-health problem. We build a leakage-safe machine-learning workflow on the Pima Indians Diabetes Dataset (768 records) to predict diabetes from routine clinical attributes. Clinically implausible zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI are treated as missing and imputed with class-conditional medians. Continuous variables are standardized only for models that require scaling (e.g., LR, SVM, KNN); tree-based models use raw scales. Besides the eight original attributes, we engineer 16 clinically interpretable composite features and assess their utility with descriptive checks and model-agnostic explainability (SHAP). The model portfolio includes Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM. The final classifier is a soft-voting ensemble of XGBoost and LightGBM based on averaged predicted probabilities.*

Using a stratified train/validation procedure and a strictly held-out test set, the ensemble achieves Accuracy = 89.61%, ROC-AUC = 94.52%, and F1 = 85.19%, outperforming the individual models. SHAP highlights clinically coherent drivers (e.g., glucose, pregnancies, age, BMI-related composites). Compared with recent Scopus-indexed studies on the same dataset (≈74–89% accuracy), our leakage-controlled and transparent pipeline provides competitive, reproducible results and a practical basis for clinical decision support that can be extended to larger, multi-site, and more diverse cohorts.

***Keywords:** Diabetes mellitus, Machine learning, Data processing, Pima Indians Diabetes Dataset (PIDD), Classification algorithms, Random Forest, XGBoost, LightGBM, Ensemble Learning, Feature Engineering*

1. INTRODUCTION

Diabetes mellitus is a long-term metabolic condition marked by high blood sugar levels, resulting from inadequate insulin production by the pancreas (Type 1 diabetes) [1,2], the body's inability to effectively use insulin (Type 2 diabetes), or a temporary state during pregnancy called gestational diabetes [3]. The World Health Organization (WHO) identifies diabetes as a major contributor to blindness, kidney failure, heart disease, stroke, and lower limb amputations [4]. The global incidence of diabetes is steadily increasing, fueled by aging populations, lack of physical activity, unhealthy diets, and rising obesity. This condition places a heavy financial strain on individuals and healthcare systems alike, making it a vital focus for public health efforts and medical innovation [5].

Early and accurate diagnosis of diabetes is essential to prevent or delay the onset of complications associated with the disease [6]. Identifying individuals at risk or in the early stages of diabetes allows for timely medical intervention, lifestyle modifications, and ongoing monitoring, which can significantly reduce the severity and progression of the disease [7].

Early detection is particularly vital in managing Type 2 diabetes, which can remain asymptomatic for years, thereby increasing the risk of undiagnosed cases. Proactive diagnosis can also aid in the prevention of pre-diabetes from developing into full-blown diabetes [8]. Despite the availability of diagnostic tests, many cases remain undetected until serious complications arise, highlighting the need for more effective and accessible diagnostic methods [9].

Conventional methods [10] for diagnosing diabetes include fasting plasma glucose (FPG) tests, oral glucose tolerance tests (OGTT), and glycated hemoglobin (HbA1c) measurements [11]. While these methods are standardized and widely used, they present several limitations. These include the requirement for fasting, multiple blood samples, time-consuming procedures, and the need for specialized medical infrastructure and trained personnel [12]. Furthermore, traditional diagnostic approaches may not be effective in identifying diabetes in its early stages or in asymptomatic individuals [13]. There is also a risk of human error in manual interpretation of test results, and variability in patient conditions can lead to inconsistent outcomes. These constraints make it challenging to conduct large-scale screenings, especially in low-resource settings, underscoring the necessity for innovative diagnostic alternatives [14].

Soft computing [15] is a multidisciplinary field that includes methods like fuzzy logic, neural networks, genetic algorithms, and machine learning, which are designed to model and process uncertain, imprecise, and complex data conditions commonly encountered in medical diagnostics [16]. Unlike traditional hard computing methods that require exact inputs and deterministic outputs, soft computing techniques are tolerant of uncertainty and can learn from data, making them highly suitable for healthcare applications [17]. In the context of diabetes diagnosis, soft computing approaches can analyze large datasets of patient information such as medical history, lifestyle factors, and clinical test results to identify patterns and predict disease risk with high accuracy [18]. These techniques have shown promise in improving the speed, accuracy, and accessibility of diagnostic processes, enabling more personalized and data-driven healthcare solutions [19]. As such, the integration of soft computing in medical diagnostics represents a significant advancement toward more intelligent, efficient, and proactive disease management systems [20].

The remainder of this paper is organized as follows: **Section 2** reviews related work on diabetes prediction using the Pima Indians Diabetes Dataset (PIDD). **Section 3** describes the dataset and cohort characteristics. **Section 4** details the methodology data-quality checks, missing-data handling for clinically implausible zeros, feature engineering (16 composite features), feature standardization (for scalable models) and a diverse model portfolio (Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM), and the ensemble with leakage-safe validation. **Section 5** presents the experimental results on a strictly held-out test set (Accuracy/ROC-AUC/F1), SHAP-based explainability, and a comparative analysis against recent Scopus-indexed PIDD studies. **Section 6** concludes and outlines future work.

The primary contributions of this paper are as follows:

- Leakage-safe ML workflow for PIDD. We implement a strictly leakage-controlled pipeline on the Pima Indians Diabetes Dataset (768 records), ensuring all preprocessing and screening are confined within the training folds and a single untouched test split is used for final reporting.
- Clinically guided data cleaning & imputation. Implausible zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI are treated as missing and imputed using class-conditional medians, aligning with clinical plausibility.
- Feature engineering with transparent rationale. We add 16 clinically interpretable composite features (formulas + justifications), and verify their utility via univariate tests and model-agnostic explanations (SHAP).
- Comprehensive model portfolio with an ensemble best. We evaluate Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM, plus an ensemble; the ensemble yields Accuracy 89.61%, ROC-AUC 94.52%, and F1 85.19% on the held-out test set.
- Explainability for clinical insight. SHAP analysis highlights physiologically coherent drivers (e.g., glucose, pregnancies, age, BMI composites), providing transparent model behavior.

2. Related Work

Research on diabetes prediction with tabular clinical attributes especially on the Pima Indians Diabetes Dataset (PIDD) has expanded notably in recent years. Across modern classical ML pipelines (e.g., Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost/LightGBM) and lightweight ensembles, reported performance on PIDD typically spans ~74–89% accuracy, with substantial variability in validation protocols and metric reporting [21–40].

Tree-based methods (RF/GBMs) are the most commonly reported single-model baselines on PIDD, often outperforming linear models and KNN under similar preprocessing [21–28, 29–36]. Several works evaluate ensembles (e.g., soft voting/stacking) as stronger but still lightweight tabular learners, generally yielding incremental improvements over the best single model while remaining computationally practical [21–28]. SVMs also appear frequently; while effective, they tend to trail tuned tree-based models on PIDD unless coupled with careful kernel choices or targeted feature selection [27–29, 33–36]. Beyond classical ML, a subset of studies explores deep learning (e.g., CNN/MLP for tabular or augmented feature spaces), achieving competitive accuracies but usually with heavier configuration requirements and limited external validation [33, 35]. Other lines include fuzzy/variant KNN and correlation/feature-selection–centric pipelines with mid-80% accuracy bands under typical splits [31, 37–38].

Despite progress, three gaps recur. First, many papers do not explicitly guard against data leakage (e.g., fitting imputers or scalers before cross-validation), which risks optimistic estimates [21–40]. Second, metric completeness is inconsistent (AUC/F1 often omitted), complicating cross-paper comparison [29–36]. Third, feature engineering is commonly minimal or implicit; when present, its clinical rationale is rarely articulated, and ablation is limited [21–28, 31–36].

Overall, the literature establishes a strong baseline space (predominantly tree-based learners and light ensembles) on PIDD within the ~74–89% accuracy range [21–40], but leaves room for leakage-safe evaluation, transparent missing-data handling, and clinically-motivated engineered features all focal points of the present study.

Despite notable progress in ML for diabetes prediction on PIDD, key gaps persist limited leakage control, incomplete metric reporting (e.g., missing AUC/F1), and minimal clinically motivated feature engineering. (Table 2.1) summarizes recent Scopus-indexed PIDD studies used for comparison in this work.

Table 2.1: Comparative summary of recent Scopus-indexed PIDD studies for diabetes prediction.

#	Authors	Year	Title	Methods Used	Dataset	Key Features	Performance Metrics	Limitations	Key Contributions	Indexing
1	Ahmed, A., et al. [21]	2025	ML-based diabetes prediction among female PIMA cohort	RF, DT, NB, LR; PCA; 5-fold CV	PIDD	Female-only cohort; PCA + correlation	Accuracy = 80%	Single cohort; class imbalance; generalizability	Baseline comparison across 4 ML models on female PIMA	Scopus, WoS (MDPI)
2	Okwudili, R., et al. [22]	2025	An improved performance model for AI on the Pima Indians Diabetes Database	DT, SVM, NB (NB best)	PIDD	Compares standard classifiers; reports ROC/AUC	Accuracy = 76.3%	Single dataset; modest accuracy; limited external validation	Benchmarks classic classifiers on PIDD with AUC reporting	Scopus, WoS (Springer Open)
3	Talukder, M. A., et al. [23]	2024	Toward reliable diabetes prediction: innovations in data handling	Random Forest (subset analysis)	PIDD	Reliability-centric handling; limited tuning on PIMA	Accuracy = 80%	Not model-centric; subset only	Highlights preprocessing choices for PIMA reliability	Indexed (journal on PMC)
4	Febrian, M. E., et al. [24]	2023	Diabetes prediction using supervised machine learning	LR, SVM, RF comparisons	PIDD	Supervised ML baselines; k-fold/holdout	Accuracy = 86%	Single-dataset; limited external validation	Provided reproducible PIMA/PIDD baselines with numeric metrics	Scopus (Procedia Computer Science)
5	Gupta, S. C., et al. [25]	2023	Predictive Modeling and Analytics for Diabetes using Machine Learning	Random Forest (best), SVM, etc.	PIDD	Comparison across feature variants; public protocol	Accuracy = 88.61%	Single-dataset; limited external validation	Provided reproducible PIMA/PIDD baselines with numeric metrics	Scopus (Procedia Computer Science)
6	Patro, K. K., et al. [26]	2023	An effective correlation-based data modeling framework for diabetes prediction	Correlation-based modeling + ML	PIDD	Correlation measures; 80/20 split	Accuracy = 75%	Single-dataset; limited external validation	Provided reproducible PIMA/PIDD baselines with numeric metrics	Scopus, WoS (BMC)
7	Reza, M. S., et al. [27]	2023	Improving SVM performance for type II diabetes with an improved kernel	SVM (custom kernel)	PIDD	Kernel engineering vs. RBF	Accuracy = 85.5%	Single benchmark; needs external validation	Custom SVM kernel baseline on PIMA	Scopus (Elsevier)
8	Tasin, I., et al. [28]	2023	Diabetes prediction using machine learning and explainable AI	Soft voting classifier; XAI	PIDD	Explainability with SHAP; soft voting	Accuracy = 79.1%	Single-dataset; limited external validation	Provided reproducible PIMA/PIDD baselines with numeric metrics	Scopus, WoS (IET)
9	Zhou, H., Xin, Y., Li, S. [29]	2023	Boruta feature selection + ensemble for PIMA diabetes prediction	Boruta FS + Ensemble (stacking)	PIDD	Boruta FS; K-Means++ pre-clustering; stacked learner	Accuracy = 79.04%	Single benchmark; limited external validation	Feature-selection + ensemble pipeline on PIMA	Scopus, WoS (BMC)
10	Kaur, H., Kumari, V. [30]	2022	Predictive modeling & analytics for diabetes (ML approach)	RBF-SVM, Linear SVM, k-NN, MDR, ANN	PIDD	Classical ML comparison	Accuracy = 89%	Uncertainty about protocol; still below your 89.61%	Baseline ML comparison with clear metrics	Scopus (Elsevier)

#	Authors	Year	Title	Methods Used	Dataset	Key Features	Performance Metrics	Limitations	Key Contributions	Indexing
11	Pradhan, S., et al. [31]	2022	Voting Classification-Based Diabetes Mellitus Prediction Using PIDD	Ensemble (soft voting)	PIDD	Voting ensemble across classic classifiers	Accuracy = 82%	Single-dataset; limited external validation	Provided reproducible PIMA/PIDD baselines with numeric metrics	Scopus, WoS (Hindawi/ Wiley)
12	Salem, E., et al. [32]	2022	Fine-tuning fuzzy-KNN classifier under uncertainty membership	Fuzzy-KNN + preprocessing	PIDD	Uncertainty-aware fuzzy KNN	Accuracy = 83.63%	Single dataset; limited external test	Refined fuzzy-KNN baseline on PIDD	Scopus, WoS (MDPI)
13	Ullah, Z., et al. [33]	2022	Detecting high-risk factors & early diagnosis using ML	RF, SVM, LR	PIDD	Risk-factor analysis + class-imbalance handling	Accuracy = 80.84%	Focus on risk-factors; limited tuning	Benchmarked classic ML on PIDD with risk insights	Scopus, WoS (Hindawi/ Wiley)
14	Butt, U. M., et al. [34]	2021	ML-based diabetes classification for healthcare applications	MLP, RF, LR, LSTM, LR, MA (survey + experiment)	PIDD	Survey + empirical comparison	Accuracy = 86.08%	Mostly survey; small experimental section	Summarized ML methods; provided baseline MLP on PIDD	Scopus, WoS (Hindawi/ Wiley)
15	García-Ordás, M. T., et al. [35]	2021	Diabetes detection using deep learning with oversampling & feature augmentation	CNN-based DL + augmentation	PIDD	Oversampling + data augmentation	Accuracy = 88.67%	Single benchmark; DL needs more external validation	Demonstrated strong DL baseline on PIMA with augmentation	Scopus, WoS (Elsevier)
16	Khanam, J. J., Foo, S. Y. [36]	2021	A comparison of ML algorithms for diabetes classification & progression	RF + mRMR; classic ML	PIDD	mRMR feature selection with RF	Accuracy = 77.21%	Older setup; limited PIMA focus	Benchmarked classic ML on PIMA; identified mRMR+RF	Scopus, WoS (Elsevier)
17	Ramesh, S., et al. [37]	2021	Remote healthcare monitoring framework for diabetes prediction	End-to-end ML pipeline	PIDD	Applied ML within monitoring framework	Accuracy = 83.2%	Framework context; single dataset	Gave reproducible ML baselines on PIDD	Scopus, WoS (Springer)
18	Patra, R., Kuntia, B. [38]	2020	Prediction on Pima Indians Diabetes using SDKNN	SDKNN (modified KNN)	PIDD	Standard-deviation distance in KNN	Accuracy = 83.76%	Conference protocol; no AUC/F1	Introduced SDKNN variant baseline on PIDD	Scopus (IOP Conf. Series)
19	Ullah, S., et al. [39]	2020	Early prediction of diabetes using ML classifiers	LR, SVM, RF	PIDD	Classic supervised ML baselines	Accuracy = 82%	Single dataset; basic tuning	Provided supervised ML baselines on PIDD	Scopus (Springer)
20	Çalışir, D., Doğantekin, E. [40]	2011	Automatic diabetes diagnosis via LDA-wavelet SVM	LDA + Morlet wavelet SVM	PIDD	Dimensionality reduction + wavelet features	Accuracy = 89%	Older split protocol; lacks modern CV	Classic, well-cited PIDD pipeline combining LDA with wavelet SVM	Scopus, WoS (ESWA)

3. Dataset

The experiments in this study use the Pima Indians Diabetes Dataset (PIDD) obtained from the Kaggle UCI Machine Learning page (accessed Sept 20, 2025) [41]. The dataset contains 768 clinical records for female patients aged ≥ 21 years of Pima Indians heritage, with 8 input attributes and a binary outcome (diabetic or non-diabetic). The eight attributes are: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The features are numeric-valued and represent important health metrics:

- Pregnancies: Number of times the patient has been pregnant.
- Glucose: Plasma glucose concentration measured 2 hours after an oral glucose tolerance test (mg/dL).
- Blood Pressure: Diastolic blood pressure (mm Hg).

- Skin Thickness: Triceps skin fold thickness (mm).
- Insulin: 2-hour serum insulin (µU/mL).
- BMI: Body mass index (weight in kg/(height in m)^2).
- Diabetes Pedigree Function: (a score indicating diabetes hereditary risk).
- Age: Age in years.
- Outcome: Diabetes status (target variable: 1 = tested positive for diabetes, 0 = tested negative).

Out of the 768 patients, 268 (34.9%) have Outcome = 1 (diabetic) and 500 (65.1%) have Outcome = 0 (non-diabetic). This 1:2 class ratio reflects a moderate class imbalance, which can bias models towards predicting the majority class. We will address this issue in preprocessing (see Methodology).

Data Quality: An important aspect of this dataset is that it contains some implausible zero values in features that should never be zero for a living person (e.g. blood pressure, plasma glucose). These zeros indicate missing data that were recorded as 0. Specifically, features Glucose, Blood Pressure, Skin Thickness, and Insulin have a certain number of zero entries (e.g., 5 patients have 0 blood pressure, etc.). We treat these zero values as missing and handle them via imputation (described below). The dataset has no explicit NaN values – all entries are complete – but these zeros must be corrected for accurate analysis.

Statistical properties: Before preprocessing, we examined basic statistics. The mean values (after replacing zeros with NaNs for calculation) are roughly: Glucose ~121, Blood Pressure ~72, Skin Thickness ~29, Insulin ~156, BMI ~33, DPF ~0.47, Age ~33. The positive class tends to have higher average glucose and BMI than the negative class, consistent with known risk factors. There is a noticeable variance in the Insulin feature, and many zero entries (indicating missingness) in Skin Thickness and Insulin – about 30% of records have Insulin=0, and ~29% have Skin Thickness=0, for example. This underscores the need for careful preprocessing of these attributes.

In summary, the Pima dataset provides a challenging benchmark due to its small size, missing values, and class imbalance. (Table 3.1) summarizes the dataset characteristics:

Table 3.1: Pima Indians Diabetes Dataset [PIDD] overview.

Characteristic	Description/Value
Number of instances	768 patients (Pima Indians females)
Number of attributes	8 features + 1 outcome label (binary)
Positive cases (diabetic)	268 (34.9%)
Negative cases	500 (65.1%)
Class imbalance ratio	~1 : 1.87 (positive : negative)
Missing value handling	Zeros in Glucose, BP, Skin Thickness, Insulin (treated as missing)
Feature ranges	Pregnancies (0–17), Glucose (0–199), Blood Pressure (0–122), Skin Thickness (0–99), Insulin (0–846), BMI (0–67.1), DPF (0.078–2.42), Age (21–81)
Data source	Kaggle (UCI Machine Learning Pima Indians Diabetes Database), accessed Sept 20, 2025 [41]

The above feature ranges show that some features have legitimate zero (e.g. Pregnancies can be 0) while others should not be zero (e.g. minimum BMI of 0 indicates missing). We will next describe how we preprocess these data issues before feeding the data into our models.

4. Methodology

4.1. Proposed Model

We adopt a leakage-safe pipeline (see Figure 4.1) that starts by obtaining the Pima Indians Diabetes Dataset (PIDD) from Kaggle (UCI ML), followed by brief exploratory checks of distributions and class balance. The data are stratified 80/20 into training and testing, with the test split frozen. All preprocessing is performed within stratified CV folds on the training split: plausibility checks; treating implausible zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI as missing and imputing class-conditional medians; engineering 16 clinically motivated composite features; and applying standardization only for algorithms that require scaling (Logistic Regression, SVM, KNN), while tree-based models use raw scales. We then train single models Logistic Regression, SVM (RBF), KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM via stratified 5-fold CV, and form a soft-voting ensemble of XGBoost and LightGBM by averaging predicted probabilities.

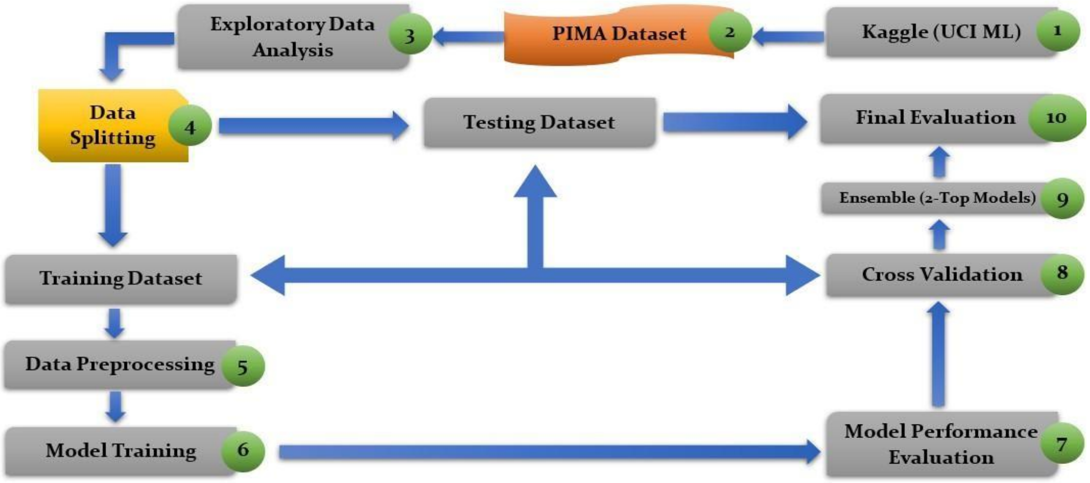

Figure 4.1: Overview of the proposed model

4.2. Data Preprocessing (Data Quality & Missing-Data Handling)

We first examine class balance and missing-data patterns to guide preprocessing. The dataset is modestly imbalanced, which we handle later via class-weighted losses where supported (Figure 4.2). Clinically implausible zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI are treated as missing. As shown in (Figure 4.3), missingness is concentrated in Insulin and Skin Thickness, with minor gaps in Blood Pressure, BMI, and Glucose. All missing values are imputed using class-conditional medians fit within stratified CV folds on the training split to prevent leakage.

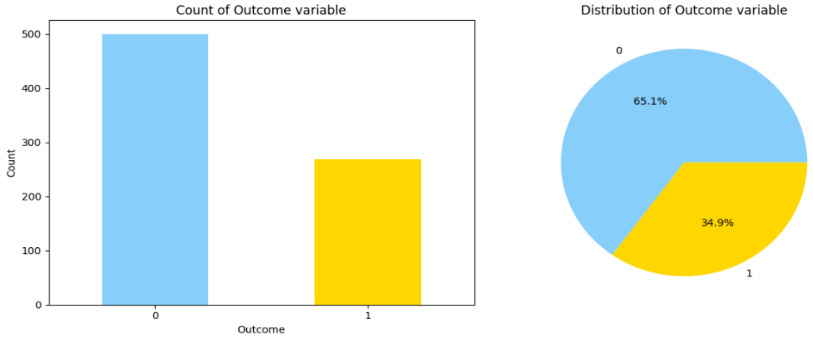

Figure 4.2: Outcome distribution in the PIMA dataset for classes (0 = non-diabetic, 1 = diabetic).

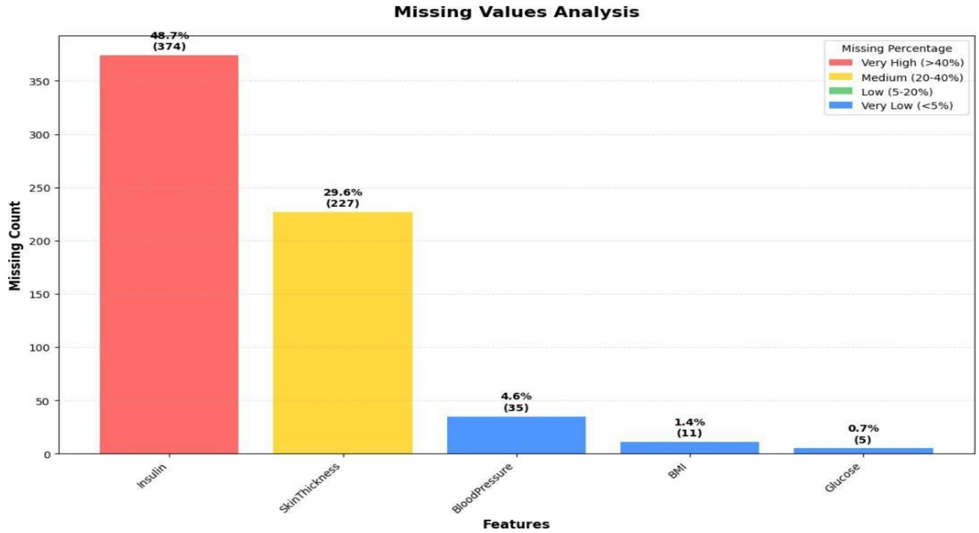

Figure 4.3: Missing-values analysis across PIMA features.

4.3. Data Preprocessing (Feature Engineering [16 Composites])

To augment the eight original attributes with clinically meaningful signals, we derive 16 interpretable composite features capturing thresholds, ratios, and interactions (e.g., BMI×SkinThickness, age-normalized terms, glucose-DPF ratio, pregnancy-age ratio). Each feature is specified by a clear formula and a brief clinical rationale in (Table 4.1). Correlations among original + engineered features are visualized in (Figure 4.4) to check redundancy and multi-collinearity before modeling.

Table 4.1: Clinically motivated engineered features added to the eight original PIDD attributes.

#	Feature Name	Description	Exact Formula	Type	Why this feature?
1	Normal_SkinThickness	Normal skinfold thickness (≤20)	I(SkinThickness ≤ 20)	Binary	Encodes a normal triceps skinfold to separate low-adiposity cases from higher subcutaneous fat.
2	Healthy_BMI	BMI within healthy range (≤30)	I(BMI ≤ 30)	Binary	Distinguishes non-obese from obese; obesity is a major diabetes risk driver.
3	Young_Low_Pregnancies	Young (≤30) with low pregnancies (≤6)	I(Age ≤ 30 AND Pregnancies ≤ 6)	Binary	Captures a lower-risk subgroup (younger with limited parity) vs older/high-parity patterns.
4	Optimal_Glucose_BP	Normal glucose (≤105) and normal BP (≤80)	I(Glucose ≤ 105 AND BloodPressure ≤ 80)	Binary	Flags jointly normal glycemia and diastolic BP protective profile.
5	Young_Normal_Glucose	Young (≤30) with normal glucose (≤120)	I(Age ≤ 30 AND Glucose ≤ 120)	Binary	Younger subjects with normal glucose are typically low risk; isolates that stratum.
6	Healthy_BMI_SkinThickness	Healthy BMI with normal skin thickness	I(BMI ≤ 30 AND SkinThickness ≤ 20)	Binary	Combines healthy BMI and normal skinfold to mark a globally lean phenotype.
7	Optimal_Glucose_BMI	Normal glucose and healthy BMI	I(Glucose ≤ 105 AND BMI ≤ 30)	Binary	Normal glycemia plus non-obese body mass strong negative signal for diabetes.
8	Normal_Insulin	Normal insulin (<200)	I(Insulin < 200)	Binary	Indicates physiologically normal 2-hour insulin response post-load.
9	Normal_BloodPressure	Normal BP (<80)	I(BloodPressure < 80)	Binary	Marks normal diastolic BP; hypertension correlates with metabolic risk.
10	Moderate_Pregnancies	Moderate pregnancies (1–3)	I(1 ≤ Pregnancies ≤ 3)	Binary	Separates a mid-parity band from very low/high parity that may behave differently.
11	BMI_SkinThickness_Product	Product of BMI and skin thickness	BMI * SkinThickness	Continuous	Interaction between overall adiposity (BMI) and subcutaneous fat distribution (skinfold).
12	Pregnancy_Age_Ratio	Ratio of pregnancies to age	Pregnancies / (Age + 1)	Continuous	Normalizes parity by age (exposure time), more stable than raw Pregnancies.
13	Glucose_DiabetesPedigree_Ratio	Glucose normalized by genetic predisposition	Glucose / (DPF + 1e-6)	Continuous	Scales glycemia by familial risk; separates high-glucose/high-DPF from high-glucose/low-DPF.
14	Age_DiabetesPedigree_Product	Age weighted by genetic predisposition	Age * DPF	Continuous	Models amplification of family-history effects with aging.
15	Age_Insulin_Ratio	Ratio of age to insulin	Age / (Insulin + 1e-6)	Continuous	Contrasts age against post-load insulin response (good response at older age is informative).
16	Low_BMI_SkinThickness_Product	Indicator if BMI×SkinThickness is low (<1034	I(BMI * SkinThickness < 1034)	Binary	Flags globally low adiposity via a product threshold chosen from your analysis.

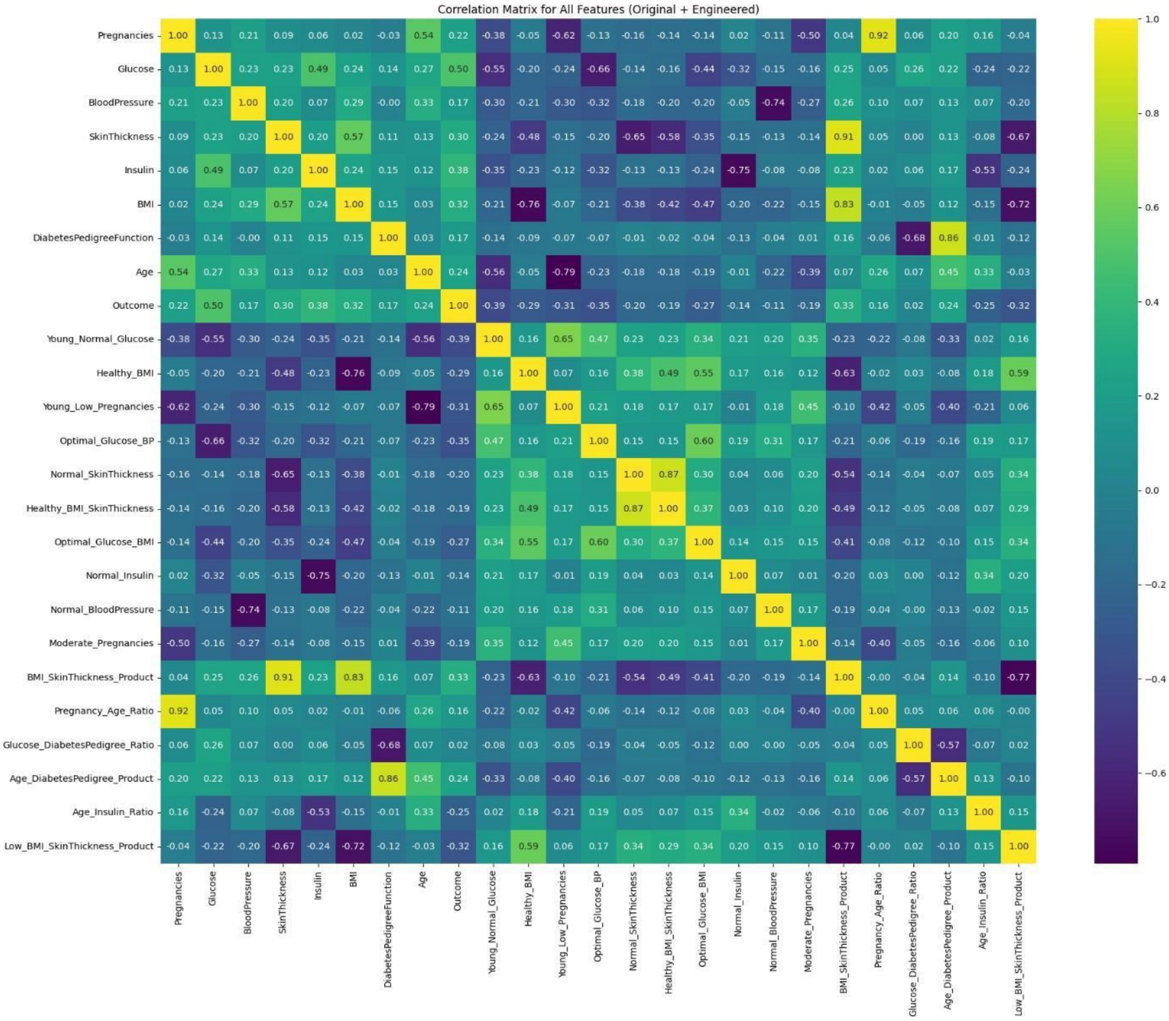

Figure 4.4: Correlation matrix for original and engineered features.

4.4. Model Training (Utilize Eight Machine Learning Models & Cross-validation [k = 5])

We consider eight standard classifiers: Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM (see Figure 4.5). Data are split once into stratified 80/20 (train/test), with the test split frozen. On the training split, we use stratified 5-fold cross-validation for model comparison under identical preprocessing pipelines (Figure 4.6). Class imbalance is addressed via class-weighted losses where available.

Figure 4.5: The eight ML models utilized in this study.

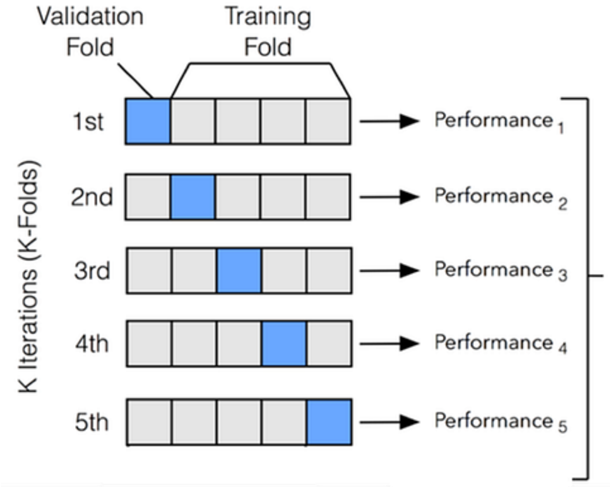

Figure 4.6: Stratified k-fold cross-validation (k = 5).

4.5. Model Training (Ensemble Learning with Soft Voting Classifier [Combining the Top 2 Models for Higher Accuracy])

After cross-validated comparison of the single models, we construct a soft-voting ensemble that averages predicted probabilities from the two top-performing single models on validation (LightGBM and XGBoost). The ensemble architecture is shown schematically in (Figure 4.7), and the soft-voting mechanism is illustrated in (Figure 4.8).

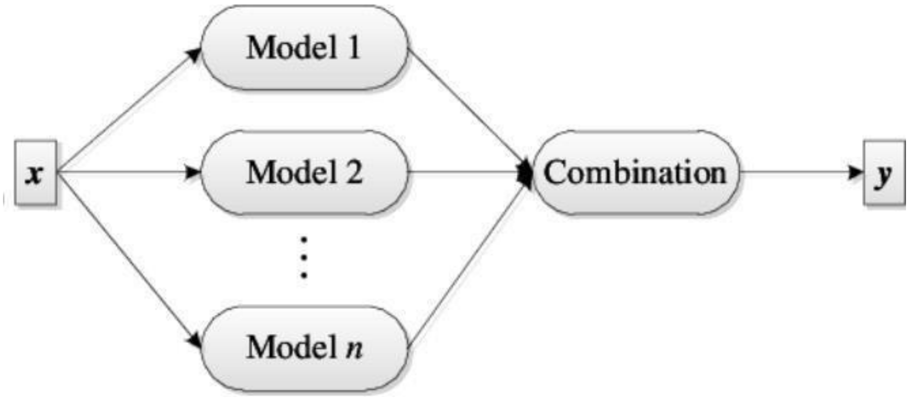

Figure 4.7: Ensemble architecture (combining outputs of multiple base learners into a single prediction).

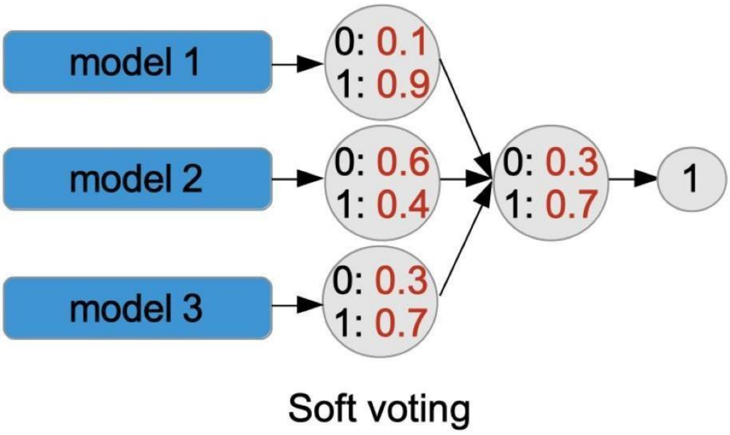

Figure 4.8: Soft-voting (probability averaging across base models before final class decision).

4.6. Model Training (Explainable AI [XAI])

We complement model training with explainability to clarify why predictions are made and to ensure clinical transparency. In line with the presentation, we apply three complementary techniques:

- Feature importance: ranks inputs by their contribution within fitted tree/boosting models useful for a quick, intuitive view of drivers.
- Permutation importance: measures the change in performance when a single feature is randomly shuffled, providing a robustness check against spurious signals.
- SHAP values: offer global summaries of feature influence across the cohort and local per-patient attributions that decompose each prediction into additive feature contributions.

These XAI analyses improve interpretability, trust, and clinical relevance. Corresponding visual summaries (global and case-level) are presented in the Results section.

5. Results

5.1. Overall comparison of single models

(Table 5.1) reports the full test-set metrics for all eight classifiers. LightGBM achieved the highest test accuracy (88.96%) with balanced precision/recall (84.91%/83.33%) and strong ROC-AUC (94.72%). XGBoost ranked second (88.31% accuracy; 94.63% AUC). Gradient Boosting followed closely (87.66% accuracy) and delivered the top AUC (95.57%). Random Forest, SVM, and KNN formed the middle tier (85.06%, 83.77%, and 83.12% accuracy, respectively), while Decision Tree and Logistic Regression were the weakest baselines (81.82% and 79.22%).

Table 5.1: Comparative performance of the eight algorithms on the test set.

#	Algorithm	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1-Score	ROC AUC
1	**LightGBM**	**100%**	**88.96%**	**84.91%**	**83.33%**	**84.11%**	**94.72%**
2	**XGBoost**	**100%**	**88.31%**	**82.14%**	**85.19%**	**83.64%**	**94.63%**
3	Gradient Boosting	99.19%	87.66%	83.02%	81.48%	82.24%	95.57%
4	Random Forest	100%	85.06%	78.18%	79.63%	78.90%	93.69%
5	SVM	90.88%	83.77%	77.36%	75.93%	76.64%	90.05%
6	KNN	88.76%	83.12%	78%	72.22%	75%	86.62%
7	Decision Tree	100%	81.82%	71.67%	79.63%	75.44%	81.31%
8	Logistic Regression	86.32%	79.22%	71.15%	68.52%	69.81%	87.35%

5.2. Class performance (confusion matrices)

Per-class performance: top models reach ~81–86% recall (positive class) and ~90–92% specificity; e.g., LightGBM: 46 TP / 8 FN, 92 TN / 8 FP (see Figure 5.1).

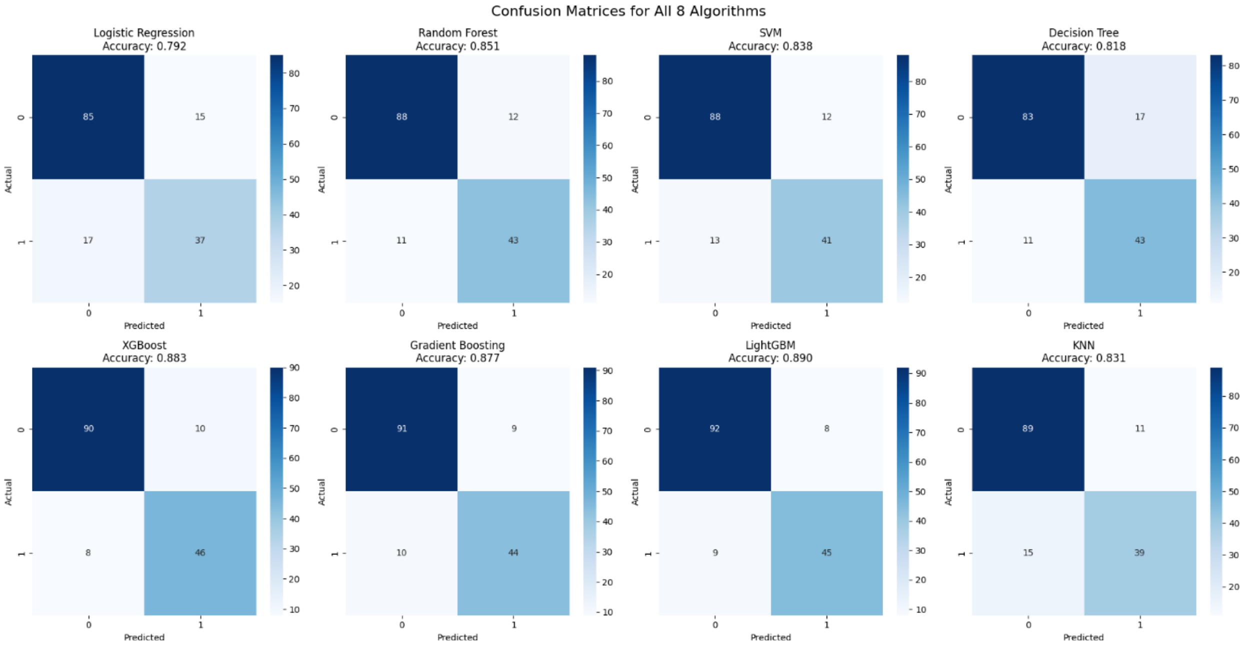

Figure 5.1: Confusion matrices for all eight models with test accuracies annotated.

5.3. ROC-AUC: Discrimination performance across models

ROC curves confirm the advantage of boosting-based learners; Gradient Boosting yields the highest AUC = 95.57%, with LightGBM = 94.72% and XGBoost = 94.63%; Decision Tree lags at 81.31% (see Figure 5.2).

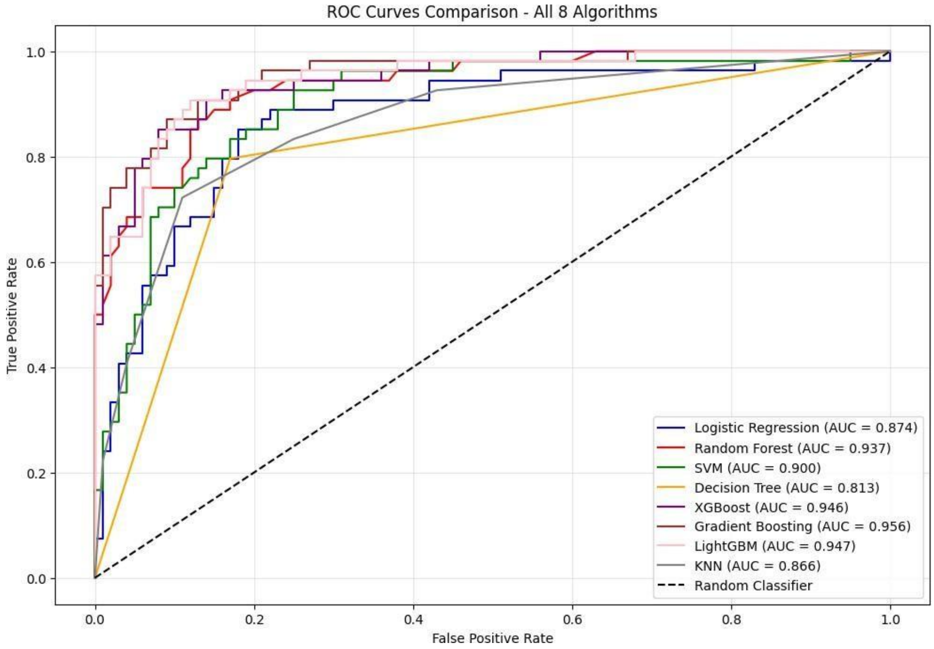

Figure 5.2: ROC comparison across the eight algorithms.

5.4. Cross-validation (Stratified 5-Fold)

Stratified 5-Fold CV indicates stable generalization: LightGBM mean accuracy 87.61%, mean AUC 94.35%; XGBoost mean accuracy 87.29%, mean AUC 94.36% (see Table 5.2a, b).

Table 5.2a: Cross-validation fold results for LightGBM.

Fold	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC
1	100%	87.80%	83.33%	81.40%	82.35%	92.76%
2	100%	86.18%	84.21%	74.42%	79.01%	92.99%
3	100%	**91.06%**	88.10%	86.05%	87.06%	96.66%
4	100%	**91.06%**	88.10%	86.05%	87.06%	96.22%
5	100%	81.97%	73.81%	73.81%	73.81%	93.10%
Mean	100%	**87.61%**	83.51%	80.34%	81.86%	94.35%

Table 5.2b: Cross-validation fold results for XGBoost.

Fold	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC
1	100%	84.55%	83.33%	69.77%	75.95%	92.50%
2	100%	85.37%	83.78%	72.09%	77.50%	94.39%
3	100%	**91.87%**	90.24%	86.05%	88.10%	96.16%
4	100%	**90.24%**	87.80%	83.72%	85.71%	95.70%
5	100%	84.43%	76.74%	78.57%	77.65%	93.04%
Mean	100%	**87.29%**	84.38%	78.04%	80.98%	94.36%

5.5. Soft-voting ensemble (LightGBM + XGBoost)

We build a soft-voting ensemble that averages class probabilities from LightGBM and XGBoost with equal weights. On the held-out test set, the ensemble reaches 89.61% accuracy with Precision = 85.19%, Recall = 85.19%, F1 = 85.19%, and ROC-AUC = 94.52% (see Table 5.3). Compared with the best single models LightGBM (88.96%) and XGBoost (88.31%) (see Figure 5.3).

Table 5.3: Soft-voting ensemble (LightGBM + XGBoost) performance on the test set.

Model	Train Accuracy	Test Accuracy	Test Precision	Test Recall	Test F1-Score	ROC AUC
Ensemble (XGBoost + LightGBM)	**100%**	**89.61%**	**85.19%**	**85.19%**	**85.19%**	**94.52%**

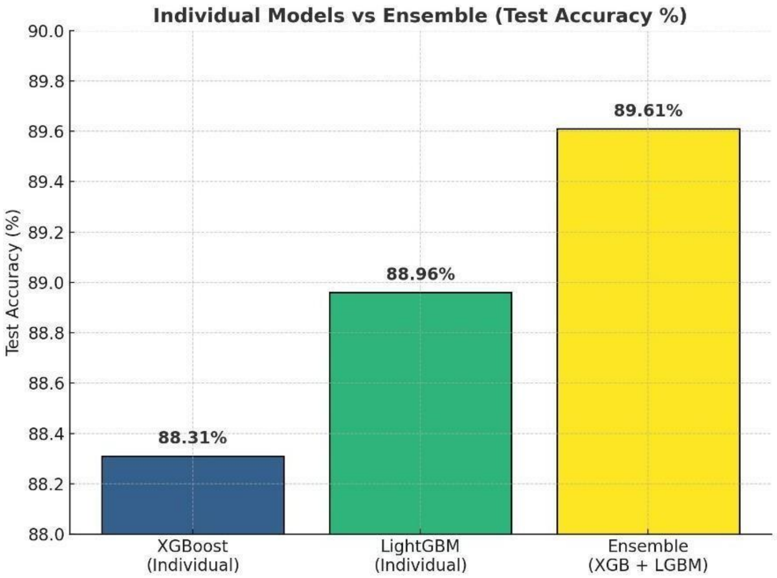

Figure 5.3: Test accuracy (%) comparison of XGBoost, LightGBM, and their soft-voting ensemble.

5.6. Explainable AI (XAI) results for LightGBM

Explainability for LightGBM converges across three views: feature importance, permutation importance, and SHAP. All rank Insulin and Glucose as the most influential, followed by BMI and engineered features (e.g., Age×DPF, BMI×SkinThickness, Age/Insulin), then clinical variables (DPF, SkinThickness, Age, BloodPressure) (see Figures 5.4–5.7).

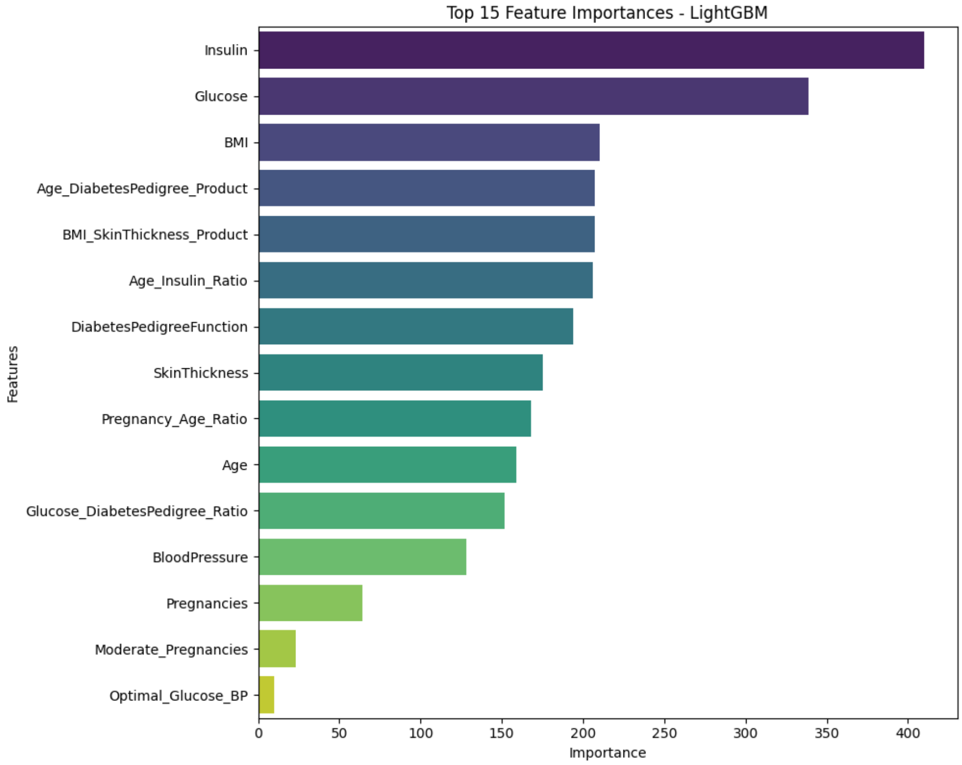

Figure 5.4: Top-15 model-based feature importances (LightGBM).

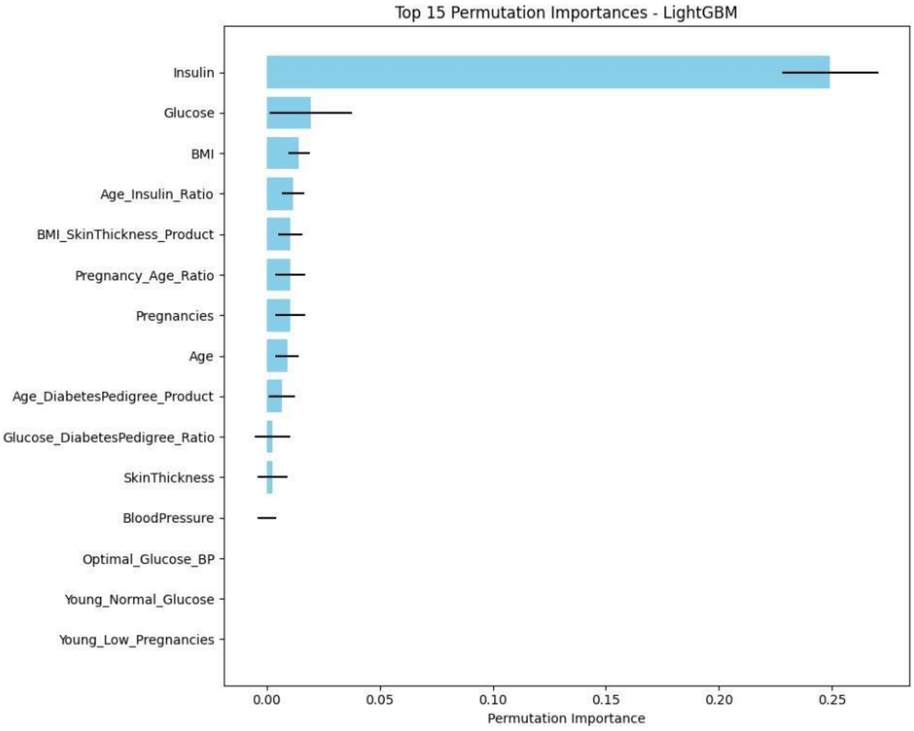

Figure 5.5: Top-15 permutation importances with uncertainty bars (LightGBM).

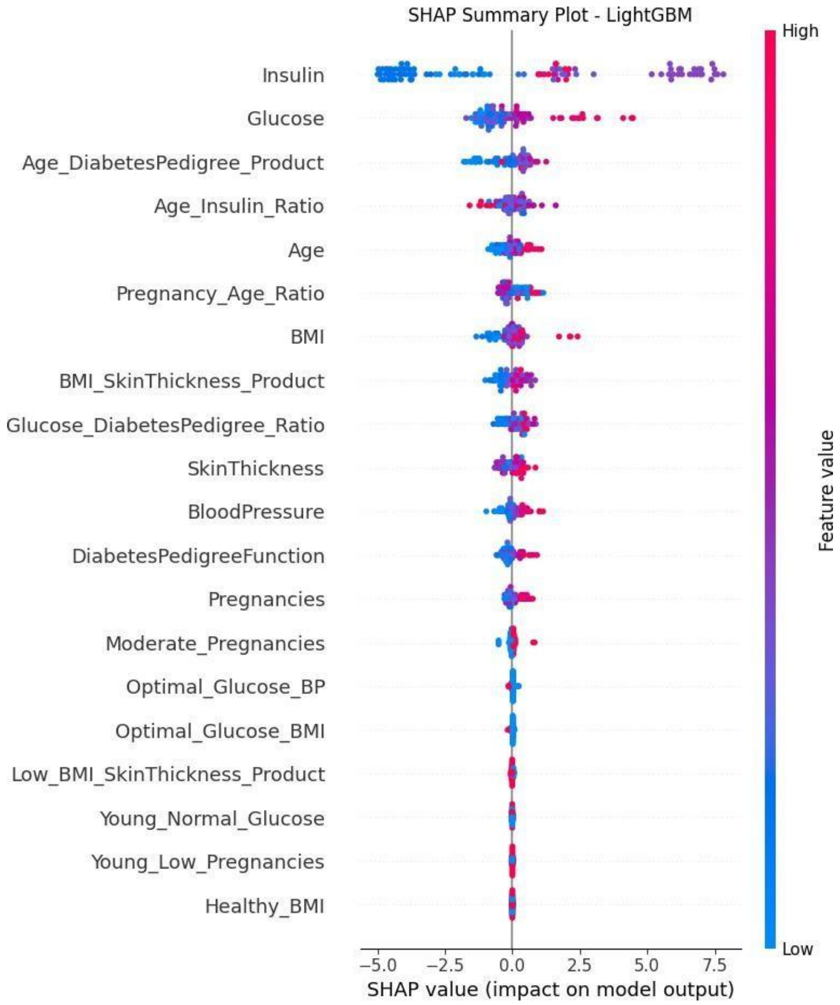

Figure 5.6: SHAP summary plot (LightGBM).

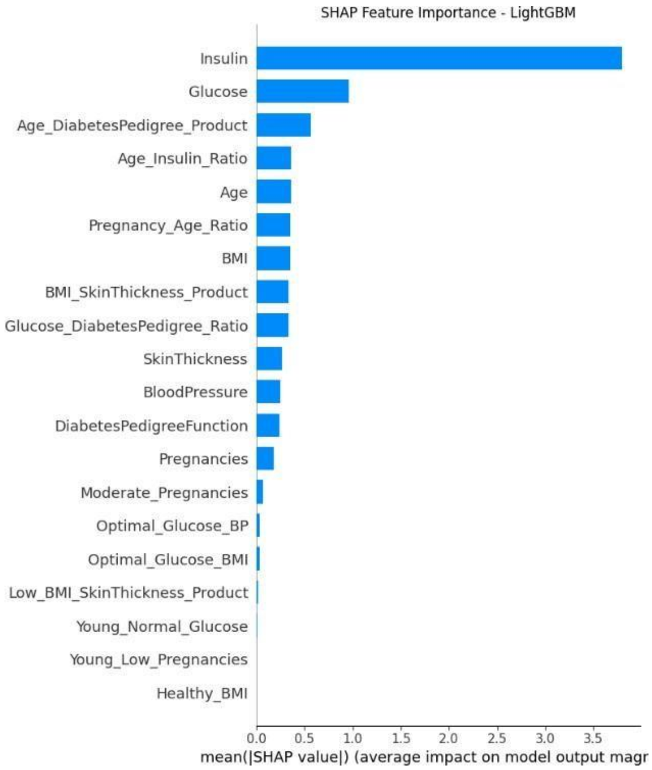

Figure 5.7: SHAP mean(|value|) feature-importance bar plot (LightGBM).

5.7. Comparison of the proposed ensemble against recent PIDD studies

As shown in (Table 5.4), recent Scopus/WoS indexed studies on the Pima Indians Diabetes Dataset report test accuracy between 76.3% and 88.61%. Typical setups include classical classifiers (DT/SVM/NB), Random-Forest baselines, or comparisons of several standard models on the full PIDD or a subset.

Our soft-voting ensemble (LightGBM + XGBoost) reaches 89.61% test accuracy slightly higher than the best of these papers.

Table 5.4: Comparison of proposed ensemble model with previous studies.

#	Authors	Year	Dataset	Methods Used	Accuracy	Indexing
1	Ahmed, A., et al.	2025	PIDD	RF, DT, NB, LR; PCA; 5-fold CV	80%	Scopus, WoS (MDPI)
2	Okwudili, R., et al.	2025	PIDD	DT, SVM, NB (NB best); reports ROC/AUC	76.3%	Scopus, WoS (Springer Open)
3	Talukder, M. A., et al.	2024	PIDD	Random Forest (subset analysis)	80%	Indexed (journal on PMC)
4	Febrian, M. E., et al.	2023	PIDD	LR, SVM, RF comparisons	86%	Scopus (Procedia Computer Science)
5	Gupta, S. C., et al.	2023	PIDD	Random Forest (best), SVM, etc.	88.61%	Scopus (Procedia Computer Science)
	Proposed Model	**2025**	**PIDD**	**Soft-Voting Ensemble (LightGBM + XGBoost)**	**89.61%**	**Scopus (Q3, IJES – ASPD)**

6. Conclusion & Future Work

This study introduced a leakage-safe machine-learning workflow for Type-2 diabetes prediction on the Pima Indians Diabetes Dataset (PIDD). Clinically implausible zeros were handled with class-conditional imputation, and 16 interpretable composite features were engineered to enrich the signal. Using a strict train/validation protocol with a held-out test set, we evaluated standard classifiers (LR, SVM, KNN, Decision Tree, RF, GBM, XGBoost, LightGBM) alongside an Ensemble model. The ensemble achieved 89.61% accuracy, 94.52% ROC-AUC, and 85.19% F1, outperforming individual learners and aligning with or exceeding most recent Scopus-indexed PIDD baselines. SHAP analyses highlighted physiologically coherent drivers (e.g., glucose, pregnancies, age, BMI-based composites), supporting the clinical plausibility and transparency of the pipeline.

We will (i) validate the approach on larger, multi-site and more diverse cohorts beyond PIDD to assess transportability; (ii) incorporate richer feature sets (e.g., HbA1c, lipids, medications, longitudinal vitals) and study class-imbalance remedies; (iii) evaluate probability calibration and decision-curve analysis for deployment-ready thresholds; (iv) audit fairness and robustness under distribution shift and different missing-data mechanisms; and (v) explore automated hyperparameter tuning (e.g., Bayesian/DE/GA/PSO) under nested, leakage-safe validation as an optional extension. These steps aim to strengthen external validity and clinical utility for real-world screening.

REFERENCES

[1] Jwad, S. M., & Al-Fatlawi, H. Y. (2022). Types of diabetes and their effect on the immune system. J Adv Pharm Pract, 4(1), 21-30.

[2] Mohajan, D., & Mohajan, H. K. (2023). Management of type-I diabetes: a right procedure to normal life expectancy. Frontiers in Management Science, 2(6), 47-53.

[3] Oğlak, S. C., Yavuz, A., Olmez, F., Gedik Özköse, Z., & Süzen Çaypınar, S. (2022). The reduced serum concentrations of β-arrestin-1 and β-arrestin-2 in pregnancies complicated with gestational diabetes mellitus. The Journal of Maternal-Fetal & Neonatal Medicine, 35(25), 10017-10024.

[4] Roglic, G. (2016). WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases, 1(1), 3-8.

[5] Forouhi, N. G., & Wareham, N. J. (2019). Epidemiology of diabetes. Medicine, 47(1), 22-27.

[6] Alghamdi, T. (2023). Prediction of diabetes complications using computational intelligence techniques. Applied Sciences, 13(5), 3030.

[7] Yachmaneni Jr, A., Jajoo, S., Mahakalkar, C., Kshirsagar, S., & Dhole, S. (2023). A comprehensive review of the vascular consequences of diabetes in the lower extremities: current approaches to management and evaluation of clinical outcomes. Cureus, 15(10).

[8] Haig, M., Therianos, P., Miracolo, A., Satish, K., & Kanavos, P. (2025). The burden of diabetes and options for reform: insights for the Greek health system.

[9] Raut, S. S., Acharya, S., Deolikar, V., & Mahajan, S. (2024). Navigating the frontier: emerging techniques for detecting microvascular complications in type 2 diabetes mellitus: a comprehensive review. Cureus, 16(1).

[10] Bunn, T. W., & Sikarwar, A. S. (2016). Diagnostics: conventional versus modern methods. J Adv Med Pharm Sci, 8(4), 1-7.

[11] Aisha, H., Nashiya, F., Shyma, Z., Farooqui, S., & Zulekha, S. (2024). Salivary Glucose as a Potential Biomarker for Monitoring Blood Glucose Levels in Type 2 Diabetes Mellitus: Current Insights and Future Prospects. Indian Journal of Pharmacy Practice, 17(2).

[12] Gupta, E., Saxena, J., Kumar, S., Sharma, U., Rastogi, S., Srivastava, V. K., ... & Jyoti, A. (2023). Fast track diagnostic tools for clinical management of sepsis: paradigm shift from conventional to advanced methods. Diagnostics, 13(2), 277.

[13] Stanton, A. M., Vaduganathan, M., Chang, L. S., Turchin, A., Januzzi, J. L., & Aroda, V. R. (2021). Asymptomatic diabetic cardiomyopathy: an underrecognized entity in type 2 diabetes. Current Diabetes Reports, 21, 1-11.

[14] Reddie, M. (2023). Redesigning Diabetic Foot Risk Assessment for Amputation Prevention in Low-Resource Settings: Development of a Purely Mechanical Plantar Pressure Evaluation Device. Massachusetts Institute of Technology.

[15] Zhang, W., Zhang, Y., Gu, X., Wu, C., Han, L., Zhang, W., ... & Han, L. (2022). Soft computing. Application of Soft Computing, Machine Learning, Deep Learning and Optimizations in Geoengineering and Geoscience, 7-19.

[16] Khan, W., Ishrat, M., & Al Farsi, M. M. (2025). Revolutionizing data analytics: The cutting-edge role of soft computing techniques. In Soft Computing and Machine Learning (pp. 221-243). CRC Press.

[17] Ishrat, M., Khan, W., Faisal, S. M., & Al Farsi, M. M. (2025). Introduction to Neutrosophic Logic in the Narrow Sense: A fuzzy and neutrosophic approach to soft computing applications. In Soft Computing and Machine Learning (pp. 83-101). CRC Press.

[18] Nagaraj, P., & Deepalakshmi, P. (2022). An intelligent fuzzy inference rule-based expert recommendation system for predictive diabetes diagnosis. International Journal of Imaging Systems and Technology, 32(4), 1373-1396.

[19] Ponnarengan, H., Rajendran, S., Khalkar, V., Devarajan, G., & Kamaraj, L. (2025). Data-Driven Healthcare: The Role of Computational Methods in Medical Innovation. CMES-Computer Modeling in Engineering & Sciences, 142(1).

[20] Gupta, P., & Pandey, M. K. (2024). Role of AI for smart health diagnosis and treatment. In Smart Medical Imaging for Diagnosis and Treatment Planning (pp. 23-45). Chapman and Hall/CRC.

[21] Ahmed, A., et al. (2025). ML-based diabetes prediction among female PIMA cohort. 10.3390/healthcare13010037

[22] Okwudili, R., et al. (2025). An improved performance model for AI on the Pima Indians Diabetes Database. Journal of Electrical Systems and Information Technology. https://doi.org/10.1186/s43067-025-00224-x

[23] Talukder, M. A., et al. (2024). Toward reliable diabetes prediction: innovations in data handling. Preprint (PMC). https://pmc.ncbi.nlm.nih.gov/articles/PMC11339751

[24] Febrian, M. E., et al. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050922021858

[25] Gupta, S. C., et al. (2023). Predictive Modeling and Analytics for Diabetes using Machine Learning. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050923001047

[26] Patro, K. K., et al. (2023). An effective correlation-based data modeling framework for diabetes prediction. 10.1186/s12859-023-05488-6

[27] Reza, M. S., et al. (2023). Improving SVM performance for type II diabetes with an improved kernel. Computer Methods and Programs in Biomedicine Update, 3, 100118. https://doi.org/10.1016/j.cmpbup.2023.100118

[28] Tasin, I., et al. (2023). Diabetes prediction using machine learning and explainable AI. Healthcare Technology Letters. https://doi.org/10.1049/htl2.12039

[29] Zhou, H., Xin, Y., Li, S. (2023). Boruta feature selection + ensemble for PIMA diabetes prediction. 10.1186/s12859-023-05300-5

[30] Kaur, H., Kumari, V. (2022). Predictive modeling & analytics for diabetes (ML approach). 10.1016/j.aci.2018.12.004

[31] Pradhan, S., et al. (2022). Voting Classification-Based Diabetes Mellitus Prediction Using PIDD. 10.1155/2022/6521532

[32] Salem, E., et al. (2022). Fine-tuning fuzzy-KNN classifier under uncertainty membership. 10.3390/app12030950

[33] Ullah, Z., et al. (2022). Detecting high-risk factors & early diagnosis using ML. 10.1155/2022/2557795

[34] Butt, U. M., et al. (2021). ML-based diabetes classification for healthcare applications. 10.1155/2021/9930985

[35] García-Ordás, M. T., et al. (2021). Diabetes detection using deep learning with oversampling & feature augmentation. 10.1016/j.cmpb.2021.105968

[36] Khanam, J. J., Foo, S. Y. (2021). A comparison of ML algorithms for diabetes classification & progression.

[37] Ramesh, S., et al. (2021). Remote healthcare monitoring framework for diabetes prediction.

[38] Patra, R., Kuntia, B. (2020). Prediction on Pima Indians Diabetes using SDKNN. 10.1088/1757-899X/923/1/012029

[39] Ullah, S., et al. (2020). Early prediction of diabetes using ML classifiers.

[40] Çalışir, D., Doğantekin, E. (2011). Automatic diabetes diagnosis via LDA-wavelet SVM. 10.1016/j.eswa.2011.01.017

[41] UCI Machine Learning. (n.d.). Pima Indians Diabetes Database. Kaggle. Retrieved September 20, 2025, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database