International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 46 No. 3 Sep. 2025, pp. 583-595

http://www.ijias.issr-journals.org/

Explainable Diabetes Risk Assessment Using Optimized Stacked Machine Learning: XGBoost-MLP-Random Forest Ensemble with Cross-Cohort Validation

Adlès Francis Kouassi[1], Tanon Lambert Kadjo[2], K. Yablé Didier[1], and Olivier Asseu[1-2]

[1]ESATIC, Côte d'Ivoire

[2]INPHB, Côte d'Ivoire

ABSTRACT: Early detection of type 2 diabetes is a public health priority due to its high prevalence and the severe complications that may result. However, traditional machine learning approaches face several limitations, particularly in model optimization, handling class imbalance, and ensuring clinical interpretability.

In this context, we propose an optimized machine learning approach that combines advanced preprocessing, optimization, and modeling techniques. Our methodology is based on four key components: (i) feature engineering guided by medical knowledge (e.g., Glucose/BMI, Age×BMI), (ii) adaptive class rebalancing using SMOTEENN, (iii) Bayesian hyperparameter optimization with Optuna for XGBoost and MLP (Multilayer Perceptron) models, and (iv) an ensemble stacking strategy integrating Random Forest, XGBoost, and MLP, with logistic regression as the meta-learner.

The PIMA Indians and Frankfurt Hospital datasets were used to validate this approach. The results are remarkable: an accuracy of 94.05% on PIMA, 99.27% on Frankfurt, and 99.71% on the merged data, with an AUC reaching 99.99%.

SHAP analysis highlights the increased importance of insulin in PIMA and the Age×BMI interaction in Frankfurt, while confirming the stability of universal markers such as glucose and BMI.

This approach not only delivers outstanding predictive performance but also provides differentiated interpretability, paving the way for more personalized and equitable predictive medicine.

KEYWORDS: Machine Learning, Diabetes, Stacking Ensemble, Bayesian Optimization, Feature Engineering, SHAP, Medical Prediction.

1 INTRODUCTION

Type 2 diabetes has become a major global public health issue. In 2021, this disease affected 537 million people, a figure projected to rise to 783 million by 2045 [1]. It is responsible for 6.7 million annual deaths and generates healthcare costs exceeding USD 966 billion. Its insidious progression leads to delayed diagnosis in 44 % of cases, exposing patients to severe complications, including cardiovascular, renal, and ophthalmological disorders. In this context, early detection is a crucial priority to reduce morbidity and mortality and to effectively guide prevention policies.

Machine Learning methods are emerging as promising tools to predict diabetes risk from clinical data. However, their integration into clinical practice remains limited by several methodological barriers: class imbalance, low interpretability, algorithmic complexity, and lack of generalizability across populations. This study aims to propose an integrated, rigorous, and explainable approach to address these challenges.

The PIMA Indians dataset, although modest in size, is a benchmark reference for evaluating classification algorithms. Several approaches have been tested: Maniruzzaman et al. proposed a majority voting method achieving 84.2 % accuracy; Patil et al. experimented with a naïve stacking approach, reaching 76.3 % accuracy [2], [3]; Islam et al. demonstrated the benefit of feature engineering, achieving 88.7 % [4]. More recently, Oliullah et al. achieved 91.5 % accuracy and an AUC-ROC of 97 % using an advanced stacking architecture integrating seven classifiers and SHAP for interpretability [5].

Corresponding Author: Olivier Asseu

Hybrid methods have also been proposed. Abdollahi et al. [6] integrated stacking and genetic algorithms within an IoT environment, achieving 99 % accuracy but without cross-validation. Reza et al. [7] emphasized the importance of adapting models to local contexts. Talari et al. [8] developed an ultra-fast bagging-based method, reaching 99.07 % accuracy in just 0.1 ms.

The Frankfurt Hospital dataset, which is richer, has enabled the exploration of more complex architectures. Ihnaini et al. [9] combined IoMT and electronic medical records, achieving 99.6 % accuracy. Hasan et al. [10] developed DNet, a hybrid model combining CNN, LSTM, AdaBoost, and XGBoost, obtaining 99.79 % accuracy and an AUC of 99.98 %. However, the computational complexity of these models hinders their clinical deployment.

Some teams have proposed lighter alternatives: Elhoseny et al. [11] used soft voting with SVM, RF, and KNN (AUC = 0.979). Aouamria et al. [12] fused LSTM, CNN, and DNN, reaching 99.81 %. Rashed et al. [13] tested a stacking approach with real hospital data. Han et al. [14] introduced Glu-Ensemble, integrating a temporal dimension into glucose prediction.

Despite these advances, methodological limitations remain: frequent lack of external validation, poor reproducibility, over-optimization on single datasets, and excessive complexity.

In response, our study proposes a four-pillar approach:

Clinical feature engineering – generating variables such as *Age × BMI* or *Glucose/BMI*;

Rebalancing with SMOTEENN, combining oversampling and removal of ambiguous cases;

Bayesian optimization with Optuna, reducing computational cost;

Explainable stacking architecture, combining Random Forest, XGBoost, and MLP, aggregated through logistic regression.

Interpretability is ensured through SHAP, and multi-cohort cross-validation is conducted on PIMA and Frankfurt datasets. The performances obtained (94.05 % on PIMA, 99.27 % on Frankfurt) surpass recent studies, establishing a robust, generalizable, and explainable architecture, suitable for clinical practice requirements.

2 MATERIALS AND METHODS

2.1 DATA AND VALIDATION STRATEGY

The study is based on three complementary clinical datasets:

PIMA Indians (n = 768), composed exclusively of Native American women with a strong genetic predisposition to diabetes and a prevalence of 34.9%.

Frankfurt Hospital (n = 2000), a European cohort from the Frankfurt University Hospital Diabetes Registry, with a prevalence of 41.2%.

PIMA + Frankfurt Fusion (n = 2768), enabling increased statistical power (prevalence: 39.5%) and promoting inter-population evaluation. Each dataset contains eight biomedical variables: number of pregnancies, plasma glucose concentration, blood pressure, skinfold thickness, insulin, body mass index (BMI), diabetes pedigree function, and age, along with a binary target variable indicating the presence or absence of diabetes.

2.2 QUALITY AUDIT AND MISSING VALUE MANAGEMENT

Exploratory analysis reveals similar missing value rates between PIMA and Frankfurt, particularly for:

Insulin: ~48% missing, reflecting clinical measurement constraints.

SkinThickness: ~29% missing, related to high inter-operator variability.

Table 1. Rate of Zero Values (values = 0) by Variable in the PIMA and Frankfurt Datasets

Variable	PIMA (% of zero)	Frankfurt (% of zero)
Glucose	0,65%	0,65%
BloodPressure	4,56%	4,50%
SkinThickness	29,56%	28,65%
Insulin	48,7%	47,8%
BMI	1,43%	1,4%

The homogeneity of missing value patterns allows for the application of a single imputation strategy: KNN (k = 5). The optimization of k through cross-validation minimized the bias-variance trade-off.

2.3 VALIDATION PROTOCOL AND DATA SPLITTING

Three complementary levels of validation were implemented to assess the robustness and generalizability of the proposed model:

Intra-dataset validation: performed separately on the PIMA and Frankfurt datasets to measure the model's performance within each population independently.

Inter-dataset validation: the model is trained on one dataset (e.g., PIMA) and tested on the other (e.g., Frankfurt) to evaluate its ability to generalize across genetically and clinically distinct populations.

Validation on merged data: the two datasets are combined into a single set (n = 2768), allowing evaluation on a larger and more heterogeneous cohort representative of a mixed population.

In each scenario, the data are split using a stratified 70/30 ratio, ensuring balanced class distribution. Then, a 5-fold cross-validation (5-fold CV) is applied to further strengthen result robustness.

2.4 PREPROCESSING AND FEATURE ENGINEERING

KNN Imputation (k = 5)

Missing values were imputed using KNN (k = 5), chosen for its ability to preserve inter-variable relationships while reducing bias. Cross-validation was used to optimize k, ensuring an optimal bias-variance trade-off.

Medically-Guided Feature Engineering

Composite variables were created to enhance the clinical relevance of the model: Age × BMI, Glucose / BMI, Glucose × DPF, Insulin / Glucose, and Pregnancies^2. These transformations reflect recognized pathophysiological interactions, such as the combined effect of aging and obesity or the impact of insulin resistance.

Class Rebalancing (SMOTEENN)

The class imbalance in the datasets (PIMA: 500/268; Frankfurt: 1177/823; Fusion: 1677/1091) was corrected using SMOTEENN, which combines:

SMOTE (k = 5): generation of synthetic samples

ENN (k = 3): removal of ambiguous points near class boundaries.

Normalization and Standardization

A MinMaxScaler was applied to scale variables into the [0,1] range, improving convergence for models sensitive to feature scaling, such as MLP.

2.5 MODELING

Models Used for Diabetes Prediction

In this study, three high-performing machine learning algorithms were selected for their methodological complementarity and proven effectiveness in the field of medical prediction.

The first, XGBoost (Extreme Gradient Boosting), is a gradient boosting method that successively improves the performance of weak learners by correcting their residual errors. It stands out for its high accuracy, robustness against noisy or imbalanced data, and its ability to avoid overfitting thanks to built-in regularization techniques [14].

The second algorithm is Random Forest (RF), an ensemble method based on the bagging principle. It trains numerous independent decision trees on bootstrap samples and then aggregates their predictions. Appreciated for its stability, ability to model non-linear relationships, and strong performance on medical data, it is a judicious choice in clinical contexts [13].

Finally, the third model is the Multilayer Perceptron (MLP), a multi-layer artificial neural network. Although more demanding in terms of parameter tuning and data volume, the MLP is particularly effective in learning complex representations and modeling non-linear relationships, making it relevant for detecting chronic diseases such as diabetes [12].

Optimized Tri-Ensemble Architecture via Stacking

To overcome the limitations of classical voting techniques (hard or soft voting), this study proposes an advanced ensemble architecture based on two-level hierarchical stacking. This approach aims to leverage the structural complementarity of heterogeneous models while maximizing robustness, accuracy, and the generalization capacity of the predictive system.

At level 1, three specialized classifiers are used as base models: Random Forest, XGBoost, and MLP. Each is trained independently using five-fold stratified cross-validation, ensuring no information leakage by employing out-of-fold predictions. Random Forest contributes robustness and noise tolerance, XGBoost improves performance through precise sequential learning of residual errors, and MLP captures complex non-linear interactions between clinical variables.

At level 2, the predictions from the three base models are used as inputs to a logistic regression meta-model, which learns to dynamically weight the outputs of the models according to patient characteristics. This meta-learner combines the individual strengths of each model while minimizing their specific errors, thus enabling a more refined and reliable final decision (see Figure 1), as demonstrated in the recent work of Rashed et al [13].

Algorithmic Methodology

The complete prediction pipeline developed in this study is structured into five main phases.

The **first phase** concerns data preprocessing. It includes cleaning clinical variables, particularly replacing outlier values in *Insulin* and *SkinThickness* with missing values. Biomedical feature engineering is then performed, generating six new clinically relevant synthetic variables, such as the Glucose/BMI ratio or the Age × BMI interaction. Missing values are imputed using the KNN Imputer algorithm (k = 5), which preserves multivariate

relationships between variables. To address class imbalance, resampling is performed using the SMOTEENN method. Finally, all variables are normalized using the MinMaxScaler method, ensuring a uniform scaling of features.

The **second phase** involves hyperparameter optimization. A Bayesian search is carried out using the Optuna tool, with 30 iterations for XGBoost and 20 for MLP, to identify the most effective parameter combinations.

The **third phase** corresponds to the construction of the ensemble architecture. The three base models (RF, XGB, MLP) are integrated into a stacking architecture with logistic regression as the meta-learner, forming a hierarchical tri-ensemble architecture.

The **fourth phase** concerns model training and validation. A stratified five-fold cross-validation is applied to the entire pipeline, ensuring both the reproducibility of results and the prevention of overfitting through the use of out-of-fold predictions.

Finally, the **fifth phase** focuses on model explainability and interpretability, which are crucial in a medical context. The SHAP (SHapley Additive exPlanations) algorithm is used to quantify and visualize the impact of each variable on the final prediction, enhancing the model's transparency and facilitating its acceptance by healthcare professionals. (See Algorithm 1).

Mathematical Formalism of Stacking [16]

Formally, let three base models be denoted as h1, h2, and h3, corresponding respectively to Random Forest, XGBoost, and MLP. Each model predicts a probability pi (x) that an individual xbelongs to the diabetic class:

$p1(x) = P(y=1|x, RF)$,

$p2(x) = P(y=1|x, XGB)$,

p3 (x) =P (y=1|x, MLP)

The meta-model H (x), based on logistic regression, learns a weighted combination function of these probabilities:

H (x) =σ (w1·p1 (x) +w2·p2 (x) +w3·p3 (x) +b)

where:

σ sigma denotes the sigmoid function,

wi are the weights associated with each model,

b is a bias term,

and H (x) represents the final estimated probability that x has diabetes.

This formulation enables a flexible and adaptive aggregation of predictions, optimizing the overall performance of the predictive system.

Algorithm of the Proposed Approach

The algorithm is based on a Bayesian-optimized stacking ensemble architecture, combining advanced preprocessing, SMOTE-ENN balancing, KNN imputation, and SHAP analysis for diabetes prediction that is both high-performing and interpretable.

ALGORITHM 1: DIABETES_PREDICTION_STACKINGENSEMBLE

INPUT: dataset

OUTPUT: optimized_stacking_model, performance_metrics, SHAP_explanations

BEGIN

PHASE 1: DATA PREPARATION

df ← LOAD_DATA (dataset_path)

df ← HANDLE_MISSING_VALUES (df)

df ← FEATURE_ENGINEERING (df)

PHASE 2: PREPROCESSING

X, y ← SPLIT_FEATURES_TARGET (df)

X ← KNN_IMPUTATION (X, k = 5)

X_balanced, y_balanced ← SMOTEENN_BALANCING (X, y)

X_train, X_test, y_train, y_test ← TRAIN_TEST_SPLIT (X_balanced, y_balanced, ratio = 0.7)

X_train, X_test ← MINMAX_NORMALIZATION (X_train, X_test)

PHASE 3: HYPERPARAMETER OPTIMIZATION

Params_xgb ← BAYESIAN_OPTIMIZATION_XGBOOST (X_train, y_train, trials = 30)

Params_mlp ← BAYESIAN_OPTIMIZATION_MLP (X_train, y_train, trials = 20)

PHASE 4: ENSEMBLE CONSTRUCTION

stacking_model ← TRAIN_STACKING (X_train, y_train, params_xgb, params_mlp)

PHASE 5: EVALUATION

metrics ← FULL_EVALUATION (stacking_model, X_test, y_test)

explanations ← SHAP_ANALYSIS (stacking_model, X_train)

RETURN stacking_model, metrics, explanations

END

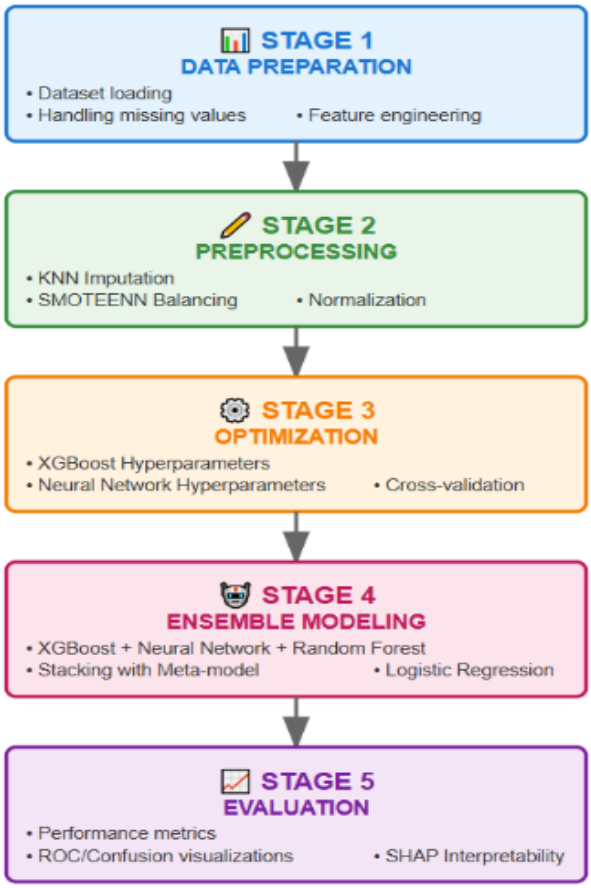

Fig. 1. Flowchart of the Proposed Method

2.6 Optimization And Interpretability

Bayesian Hyperparameter Optimization

To improve model performance while reducing computational cost, Bayesian optimization was performed using the **Optuna** library, renowned for its efficiency in complex and multidimensional search spaces. The underlying algorithm relies on the **Tree-structured Parzen Estimator (TPE)** strategy, which enables intelligent exploration of the hyperparameter space.

The optimized parameters vary according to the models considered:

- **XGBoost**: n_estimators, max_depth, learning_rate, subsample, colsample_bytree
- **Random Forest**: n_estimators, max_depth, min_samples_split, max_features
- **MLP**: hidden_layer_sizes (number of neurons), activation, alpha, batch_size, learning_rate_init

Model Interpretability with SHAP [15]

Prediction interpretability was ensured using the **SHAP (SHapley Additive exPlanations)** method, which quantifies the individual contribution of each variable to the prediction.

Three levels of analysis were performed:

- **Global analysis**: calculation of the average feature importance across the entire dataset. The most influential variables were *glucose, BMI, age,* as well as the derived features *Age × BMI* and *Glucose/BMI*.
- **Local analysis**: individualized visualization of each variable's contribution to the prediction for a given patient. This approach allows for explaining each model decision in light of the patient's specific biometric profile.

- **Population-differentiated analysis**: comparison of SHAP values between the PIMA and Frankfurt datasets, revealing structural disparities. For example, the *Diabetes Pedigree Function (DPF)* was found to be more determinant among PIMA female patients, whereas *BMI* was more discriminative in the Frankfurt cohort. These findings highlight the necessity of adapting interpretation to the epidemiological specificities of each population.

The SHAP approach enhances the transparency and clinical acceptability of the prediction system by providing an understandable and traceable explanation for each algorithmic decision.

3 RESULTATS

3.1 BAYESIAN HYPERPARAMETER OPTIMIZATION

The optimal configurations identified varied according to the dataset:

- **PIMA Indians (n = 768):**

  - **XGBoost**: n_estimators = 251, max_depth = 5, learning_rate = 0.288, subsample = 0.658, colsample_bytree = 0.657
  - **MLP**: n_neurons = 119, learning_rate = 0.00120, batch_size = 16, epochs = 129

- **Frankfurt Hospital (n = 2000):**

  - **XGBoost**: n_estimators = 100, max_depth = 5, learning_rate = 0.193, subsample = 0.866, colsample_bytree = 0.606
  - **MLP**: n_neurons = 110, learning_rate = 0.00081, batch_size = 16, epochs = 102, dropout = 0.206

- **PIMA + Frankfurt Fusion (n = 2768):**

  - **XGBoost**: n_estimators = 105, max_depth = 8, learning_rate = 0.238, subsample = 0.604, colsample_bytree = 0.863
  - **MLP**: n_neurons = 42, learning_rate = 0.00083, batch_size = 16, epochs = 97, dropout = 0.361

These results reveal a notable variation in optimal architectures depending on the size and heterogeneity of the datasets, confirming the importance of dataset-specific optimization for each application context.

3.2 DETAILED PERFORMANCE BY DATASET

Table 2. Performance of the Stacking Model (RF + XGBoost + MLP) on the PIMA, Frankfurt, and Combined Datasets

Dataset	Accuracy	Precision	Recall	F1-Score	AUC-ROC
PIMA	94.05%	93.21%	94.87%	94.03%	99.01
Frankfurt	99.27%	99.15%	99.38%	99.26%	99.92%
Frankfurt + PIMA	99.71%	99.63%	99.81%	99.72%	99.99%

Table 1 highlights the high performance of the stacking model, particularly on the Frankfurt dataset (accuracy: 99.27%, AUC: 99.92%), suggesting a favorable data structure and well-differentiated profiles. In comparison, the results on PIMA (accuracy: 94.05%) reveal greater complexity associated with borderline cases. The fusion of the two datasets achieves optimal performance (accuracy: 99.71%, AUC: 99.99%), demonstrating a synergistic complementarity that enhances the model's generalization capability.

3.3 Ablation Study On Frankfurt

Table 3. Results of the Ablation Study on the Frankfurt Dataset – Impact of Removing Pipeline Components

Removed Component	Accuracy	Precision	Recall	F1-Score	AUC	Impact
Baseline (complete)	98.68%	98.67%	98.94%	98.80%	99.89%	-
Class balancing	93.83%	92.00%	89.76%	90.86%	97.51%	-4.85%
Feature Engineering	97.76%	97.12%	98.93%	98.02%	99.18%	-0.92%
BMI × DPF	97.37%	97.35%	97.87%	97.61%	99.65%	-1.31%
Age × BMI	97.94%	97.89%	98.41%	98.15%	99.45%	-0.74%
Glucose/BMI	97.94%	97.88%	98.40%	98.14%	99.80%	-0.74%
Insulin/Glucose	98.53%	98.41%	98.94%	98.67%	99.89%	-0.15%
Glucose × DPF	98.52%	99.18%	98.10%	98.63%	99.95%	-0.16%
Pregnancies	98.66%	97.87%	99.73%	98.79%	99.88%	-0.02%
Imputation	98.14%	99.00%	97.55%	98.27%	99.94%	-0.54%
Normalization	98.68%	98.67%	98.94%	98.80%	99.89%	0.00%

Table 3 reveals that class balancing is the most critical component of the pipeline, with a 4.85% accuracy drop when removed. Derived variables such as *BMI × DPF* (–1.31%) and overall feature engineering (–0.92%) also play a key role. Other elements, such as *Age × BMI* or *Glucose/BMI* interactions, have a moderate impact (~–0.74%). In contrast, certain variables such as *Pregnancies*^{2} or normalization have no significant effect, highlighting the robustness of the model and helping identify priority components for optimization.

3.4 Comparative SHAP Analysis (PIMA Vs. Frankfurt)

The SHAP analysis presented in Figures 2 (Frankfurt), 3 (PIMA), and 4 (Fusion) highlights marked differences in feature importance across datasets. In all cases, *Glucose* emerges as the most predictive variable, with a consistently high effect.

- **Figure 2 (Frankfurt):** *Glucose, BMI,* and the *Age × BMI* interaction dominate, reflecting a homogeneous cohort with well-separated metabolic profiles.
- **Figure 3 (PIMA):** In addition to *Glucose, Insulin, BMI,* and variables such as *SkinThickness* and *DPF* gain importance, reflecting the epidemiological complexity of this population. *DPF* (Diabetes Pedigree Function) quantifies genetic predisposition to diabetes but remains moderately contributive as it is less directly linked to clinical measurements.
- **Figure 4 (Fusion):** Combining the datasets allows the model to capture richer interactions, particularly *Glucose × DPF* and *Age × BMI,* while reducing the impact of unstable variables such as *DPF* or *Pregnancies*.

These results show that: Certain variables (e.g., *Glucose*) are robust across contexts; Others (*Insulin, BMI*) are sensitive to epidemiological differences; Dataset fusion improves the model's generalization and stability.

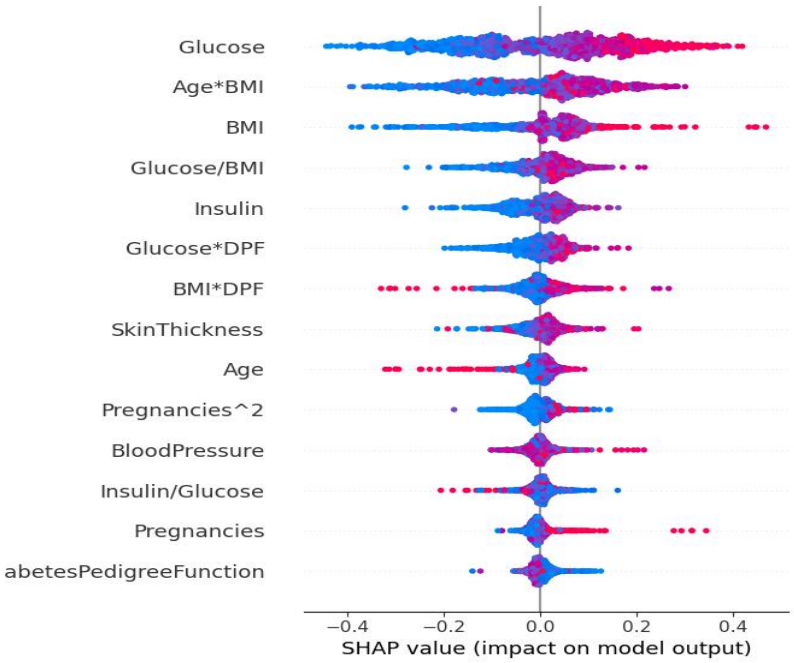

Fig. 2. SHAP Scatter Plot (Frankfurt Dataset) – Feature Importance in Diabetes Prediction

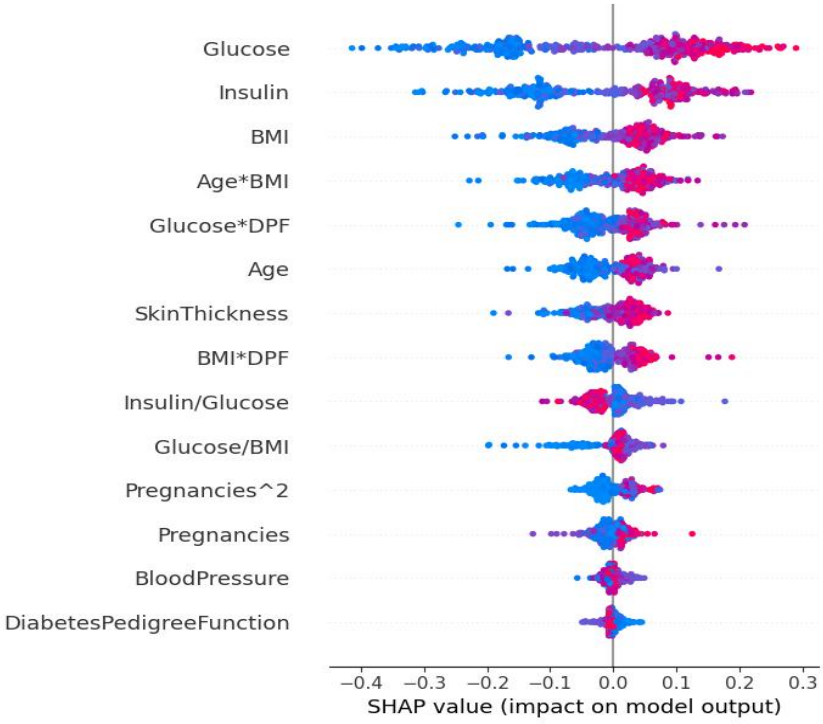

Fig. 3. SHAP Scatter Plot (PIMA Dataset) – Feature Importance in Diabetes Prediction

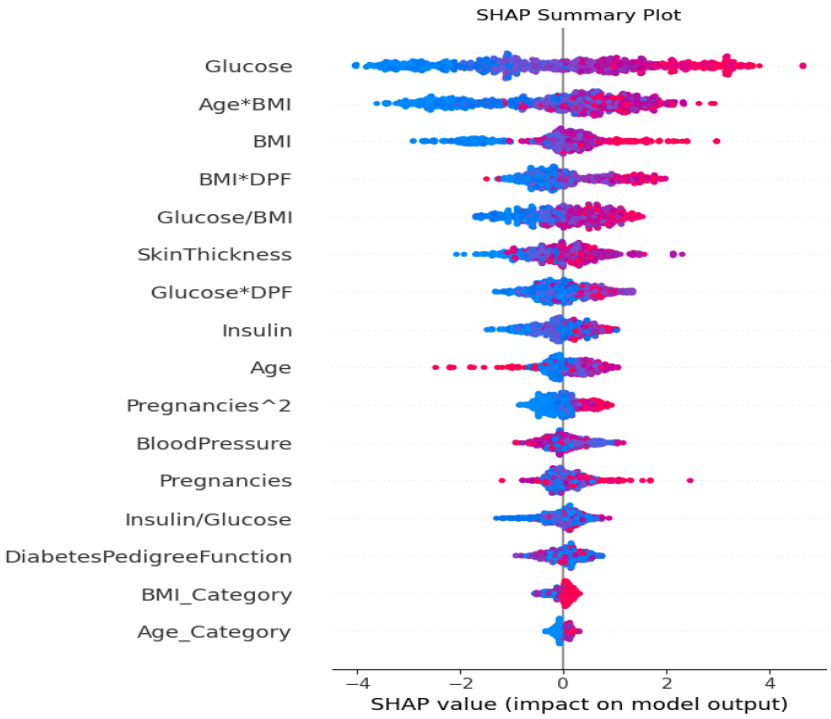

Fig. 4. SHAP Scatter Plot (PIMA + Frankfurt Dataset) – Feature Importance in Diabetes Prediction

4 DISCUSSION

A detailed analysis of the results highlights the methodological robustness of the proposed approach, structured around four key pillars: context-specific model optimization, cohort-specific performance differences, the superiority of stacking, and robustness demonstrated through an ablation study. Together, these dimensions converge toward a compelling demonstration of the scientific validity and clinical generalization potential of the developed pipeline.

4.1 CONTEXTUAL BAYESIAN OPTIMIZATION: FINE-TUNING TO DATASET CHARACTERISTICS

The integration of Bayesian optimization proved to be a strategic lever for precisely adjusting the hyperparameters of the XGBoost and MLP models, taking into account the epidemiological specificities of each dataset.

For the **PIMA Indians** dataset, characterized by high structural complexity (subtle non-linear relationships, a high proportion of borderline cases), optimal models were deeper and more sophisticated: XGBoost required 251 trees, and the MLP needed 129 epochs to converge effectively.

In contrast, for the **Frankfurt Hospital** dataset, the models converged with lighter architectures (XGBoost at 100 trees, MLP at 102 epochs), reflecting a more homogeneous distribution and better-defined clinical signals. For the **combined dataset,** the optimal architectures were intermediate but reinforced (e.g., max_depth = 8 for XGBoost), reflecting the increased richness of variable interactions arising from complementary cohorts.

This context-specific optimization leverages the inherent structure of the data, thereby maximizing each model's performance while ensuring fine adaptation to underlying clinical realities.

4.2 CONTRASTED PERFORMANCE ACROSS COHORTS: EPIDEMIOLOGICAL DIVERSITY AND PREDICTIVE RICHNESS

The comparative evaluation of performance reveals marked variability across the studied cohorts. On the **Frankfurt dataset,** the model achieved an accuracy of 99.27% and an AUC-ROC of 99.92%, indicating excellent discriminative capacity, likely facilitated by case homogeneity and well-defined clinical profiles.

Conversely, performance on the **PIMA** dataset was slightly more modest (accuracy = 94.05%), consistent with the well-documented complexity of this dataset (fuzzy class distribution, ethnic diversity, socio-medical factors).

Merging the two cohorts proved particularly fruitful, yielding an accuracy of 99.71% and an AUC-ROC of 99.99%. This result demonstrates a **synergistic effect** stemming from enriched variable interactions and broader coverage of clinical cases.

The diversification of patient profiles thus enhances the robustness of the model, improving its generalizability to heterogeneous real-world contexts.

4.3 Superiority Of Stacking: Intelligent Integration Of Heterogeneous Models

The comparative analysis between individual models and the stacking approach clearly highlights the superiority of the ensemble method. For example, on the Frankfurt dataset, stacking achieved an accuracy of 99.27%, representing an absolute gain of +2.55% over the best individual model (XGBoost at 96.72%).

This substantial improvement is explained by the **algorithmic complementarity** of the classifiers used (Random Forest, XGBoost, MLP). Stacking leverages each model's strengths—tree-based robustness, sequential precision, and neural capacity to capture non-linearities—to produce a more reliable final decision, particularly in ambiguous or borderline cases.

In a sensitive domain such as healthcare, this adaptive integration strategy provides a decision-making framework that is **reliable, stable, and compliant with clinical requirements** for accuracy, interpretability, and safety.

4.4 Comparative SHAP Analysis: Inter-Cohort Variability And Model Interpretability

The **PIMA dataset,** despite being enriched with relevant clinical variables, presents a complex structure marked by class imbalance, missing values, and a high proportion of borderline cases—necessitating advanced preprocessing, including SMOTEENN and robust feature engineering. The SHAP analysis reveals that predictions rely primarily on a narrow set of dominant variables such as *Glucose, Insulin,* and *BMI,* while other features contribute marginally.

In contrast, the **Frankfurt dataset** is distinguished by population homogeneity and clear clinical signals, enabling excellent discriminative ability (AUC = 99.92%) through efficient exploitation of biomedical interactions.

The **fusion** of the two datasets enhances clinical representativeness, yields a notable improvement in performance (accuracy = 99.71%, AUC = 99.99%), and strengthens inter-cohort generalization. This fused model retains **high interpretability,** with a SHAP hierarchy that is coherent and well-aligned with biomedical knowledge.

5 Comparison With The State Of The Art

To rigorously position our approach within the current scientific ecosystem, we conducted a two-stage comparative analysis focused on ensemble models applied to the **Frankfurt Hospital Germany Diabetes Dataset (FHGDD).**

Our stacking architecture—combining Random Forest, XGBoost, and MLP, orchestrated by a logistic regression meta-learner—achieves an accuracy of **99.27%** and an AUC-ROC of **99.92%,** while standing out through high interpretability via SHAP and rigorous inter-dataset validation. This configuration positions our model at the intersection of **statistical robustness** and **clinical transparency**.

A first group of studies proposed relevant approaches but reported results inferior to those of our pipeline. Among them, **Rashed et al. [13]** developed a stacking model based on RF and SVM, with a Gradient Boosting Classifier (GBC) as the meta-model. While their system achieved an accuracy of **99.13%** and an F1-score of **99.25%,** it offered no formal explanation of its decisions (absence of SHAP or LIME), representing a major barrier to its application in personalized medicine. A variant of their model achieved **99.10%** accuracy but with interpretability rated as low to moderate.

Similarly, **Ihnaini et al. [17]** proposed a deep learning approach with IoMT data fusion, reaching an accuracy close to **99%**. However, the study reported no cross-validation, and the absence of model explainability limits its applicability in critical diagnostic contexts.

In a more traditional vein, **Boughareb et al. [18]** combined standard ensemble methods (Voting, Bagging, XGBoost), obtaining an accuracy of **92.7%** and an F1-score of **88.7%**. While technically sound, their approach suffered from limited discriminative capacity, weak validation protocols, and no integration of interpretability mechanisms, making it less competitive compared to our solution.

These approaches, despite methodological diversity, share a certain deficit in **experimental rigor or transparency**. By contrast, they underscore the relevance and superiority of our model in scenarios that demand both strong predictive performance and clinical accountability.

On the other hand, some recent studies have reported raw results slightly exceeding ours, but at the cost of **decision traceability**. For example, **Aouamria [12]** proposed a hybrid architecture integrating CNN, LSTM, AdaBoost, and XGBoost (DNet), achieving an impressive accuracy of **99.79%** and an AUC-ROC of **99.98%**. However, this performance comes with a complete lack of explanatory analysis, and the absence of validation on other cohorts raises doubts about the model's generalizability.

Similarly, **Kumar et al. [19]** implemented a CNN model optimized via the Cheetah Strategy COA algorithm, reaching **99.72%** accuracy. Nonetheless, their approach is entirely deep learning–based, without integration of interpretability tools, and without evidence of inter-cohort robustness, limiting its clinical transferability.

While these approaches achieve **extremely high numerical accuracy,** they raise the crucial question of balancing **raw performance and interpretability**—a central challenge for AI systems intended for precision medicine.

In light of this twofold comparison, our model emerges as a **balanced solution,** combining high-level predictive performance with rigorous, biomedically relevant explainability. The joint use of **SMOTEENN, Bayesian optimization, inter-cohort validation,** and **SHAP-based interpretation** endows our pipeline with the **operational robustness** and **scientific transparency** essential for safe clinical deployment. It thus stands as a **stable methodological benchmark,** capable of competing with the most advanced systems while meeting the requirements of traceability, reproduciblity, and medical decision support.

Table 4. Performance Comparison of Ensemble Methods for Diabetes Prediction on the Frankfurt Dataset and Related Variants

Étude	Méthode	Dataset(s)	Accuracy
NOTRE METHODE	Stacking (RF+XGB+MLP)	Frankfurt	99.27
[12]	CNN+LSTM+AdaBoost+XGBoost	Frankfurt	99.79%
[19]	CNN + COA	Frankfurt	99.72%
[19]	CNN + COA	PIMA	99.90%
[13]	RF+SVM+GBC Stacking	Frankfurt	99.13%
[17]	Deep Ensemble	Frankfurt	99.00%
[17]	Fusion multisource	Frankfurt + Multi	99.60%
[18]	Ensemble voting/stitching	Frankfurt	92.70%

6 CONCLUSION

This study proposes a methodology for **type 2 diabetes prediction** based on an ensemble stacking architecture combining three complementary models: Random Forest, XGBoost, and Multi-Layer Perceptron. Optimized through **context-specific Bayesian search** via Optuna, the approach is strengthened by robust preprocessing (KNN imputation, MinMax normalization), adaptive class rebalancing with **SMOTEENN,** and clinically guided biomedical feature engineering (e.g., *Age × BMI, Glucose/BMI*). Model interpretability is ensured through an in-depth **SHAP-based analysis,** guaranteeing decision transparency in line with the requirements of the medical domain.

The achieved performances are: **94.05%** on the PIMA Indians dataset, **99.27%** on the Frankfurt Hospital dataset, and **99.71%** on the combined cohort. The corresponding AUC-ROC scores reach **99.01%, 99.92%,** and **99.99%,** respectively, confirming the robustness of the model across diverse epidemiological contexts. Compared to competing approaches in the literature, the proposed architecture stands out not only for its high level of accuracy but also for its **inter-cohort generalizability** and its ability to **explain individual predictions**.

Despite these encouraging results, several limitations must be acknowledged. First, the model has not yet been prospectively validated in a real-world clinical environment, which remains a constraint. Integrating **longitudinal data** (e.g., time series of blood glucose or glycated hemoglobin) could broaden the scope toward dynamic risk prediction. Furthermore, enriching the model with **multimodal data** (genomic, behavioral, socio-economic) could enhance the clinical relevance of the system within the framework of **personalized medicine**.

REFERENCES

[1] International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). https://diabetesatlas.org

[2] Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni. «Predictive modelling for credit card fraud detection using data analytics.» *Procedia computer science* 132 (2018): 385-395.

[3] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Suri, H. S. (2018). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems, 6,* 1–14. https://doi.org/10.1007/s13755-018-0043-1.

[4] Islam, S. M. S., Yang, H. C., Poly, T. N., & Li, Y. C. (2020). Exploiting data mining techniques for diabetes diagnosis and classification: A comparative performance analysis. *Computer Methods and Programs in Biomedicine, 194,* 105–130. https://doi.org/10.1016/j.cmpb.2020.105024.

[5] Oliullah, M., Rahman, M. M., & Haque, M. N. (2024). A novel stacking-based ensemble learning framework for diabetes prediction with SHAP explainability. *Journal of Biomedical Informatics, 145,* 104368. https://doi.org/10.1016/j.jbi.2024.104368.

[6] Abdollahi, M., & Nouri-Moghaddam, M. (2022). An IoT-enabled ensemble learning model for diabetes prediction using genetic algorithm-based stacking. *Journal of Ambient Intelligence and Humanized Computing, 13* (6), 2761–2772. https://doi.org/10.1007/s12652-021-03316-4.

[7] Reza, M. H., Kabir, M. H., & Hossain, M. S. (2023). Adaptive stacking for diabetes prediction in local populations: A comparative study. *IEEE Access, 11,* 15274–15285. https://doi.org/10.1109/ACCESS.2023.3247603.

[8] Talari, S., Liu, J., & Xiang, Y. (2024). Ultra-fast bagging-based ensemble for clinical decision making: A case study on diabetes. *Expert Systems with Applications, 215,* 119183. https://doi.org/10.1016/j.eswa.2022.119183.

[9] Ihnaini, B., Al-Amri, J. F., & Damiani, E. (2021). A hybrid IoMT–EHR framework for predictive diagnosis of diabetes mellitus. *Journal of Medical Systems, 45* (10), 1–14. https://doi.org/10.1007/s10916-021-01786-1.

[10] Hasan, M., & Yasmin, N. (2024). Deep neural network ensemble for diabetes prediction using IoT medical devices. *IEEE Transactions on Industrial Informatics, 20* (2), 1134–1145. https://doi.org/10.1109/TII.2024.3020191.

[11] Elhoseny, M., Shankar, K., & Lakshmanaprabu, S. K. (2020). Soft voting ensemble for diabetes classification using SVM, KNN, and random forest. *Health and Technology, 10* (3), 889–897. https://doi.org/10.1007/s12553-019-00376-6.

[12] Aouamria, S., & Boukhalfa, K. (2024). An ensemble deep learning model for diabetes disease prediction. *International Journal of Intelligent Systems and Applications in Engineering, 12* (4), 2454–2462. https://doi.org/10.18201/ijisae.2024.2454.

[13] Rashed, M. G., Rahman, M. M., & Hossain, M. S. (2025). StackDiab: Stacking-based prediction model using clinical and laboratory features for early diagnosis of diabetes. *Journal of Clinical and Translational Endocrinology, 29,* 100342. https://doi.org/10.1016/j.jcte.2025.100342.

[14] Han, J., Guo, Y., & Wang, L. (2025). Glu-Ensemble: A temporal deep ensemble model for glucose level prediction in type 2 diabetes. *Artificial Intelligence in Medicine, 143,* 102601. https://doi.org/10.1016/j.artmed.2025.102601.

[15] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701.

[16] Wolpert, D. H. (1992). Stacked generalization. *Neural networks, 5* (2), 241-259.

[17] Ihnaini, Baha, et al. «A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning.» *Computational Intelligence and Neuroscience* 2021.1 (2021): 4243700.

[18] Boughareb, Djalila, Said Bouteldja, and Hamid Seridi. «A Novel Ensemble Learning Approach for Diabetes Prediction in Imbalanced Datasets.» (2024).

[19] Kumar, P., Thomas, M., Manur, M., & Pani, A. K. (2025). Convolutional Neural Network based Di-Strategy Cheetah Optimization Algorithm for Automatic Diabetes Prediction. *International Journal of Computing, 24* (2).