Bridging the Generalizability Gap in Diabetes Prediction: A Noise-Resilient Hybrid Imputation-Ensemble (NR-HIE) Framework

Silicone Global Tech Gilgit Baltistan

Hasnain Alam

alamhasnain457@gmail.com

+92-355-4286542

Department: Artificial Intelligence and Data Science

Supervisor: Hafizuddin

Table of Contents

Table of Figures

Table of Tables

Abstract

The rapid escalation of Diabetes Mellitus as a global health crisis necessitates the development of robust automated diagnostic tools. While Machine Learning (ML) models have demonstrated high efficacy in controlled environments, their deployment in clinical settings is often hampered by the "Generalizability Gap" a failure to maintain performance across diverse patient demographics and varying data quality. This study addresses this critical failure mode, specifically targeting the challenges posed by Missing Not At Random (MNAR) data and dataset shift. We propose and implement the **Noise-Resilient Hybrid Imputation–Ensemble (NR-HIE) Framework**, a novel architecture designed to enhance diagnostic stability. The methodology employs a triple-stream hybrid imputation strategy to reconstruct missing clinical markers without introducing bias. This integrates **Multivariate Imputation by Chained Equations (MICE)** to capture global correlations, **K-Nearest Neighbors (KNN)** for local structural preservation, and **Robust Statistics** to mitigate the influence of outliers. These streams feed into a Stacked Generalization ensemble comprising Logistic Regression, Random Forest, and XGBoost as base learners, optimized via hyperparameter tuning. The frameworks were validated internally on the Pima Indians Diabetes dataset and externally on the DiaBD (Bangladesh) cohort, with strict protocols to prevent data leakage. Results indicate a statistically significant improvement in robustness. While the optimized Random Forest and Stacking Ensemble achieved identical internal accuracy (**77.27%**), the NR-HIE Stacking Ensemble demonstrated superior generalizability on the external cohort. It achieved an accuracy of **81.62%**, outperforming the single best model (**80.26%**). Furthermore, the Stacking Ensemble exhibited superior calibration with a **Brier Score of 0.1273**, compared to 0.1302 for the single best model. Statistical significance was confirmed via the Wilcoxon Signed-Rank Test ($p < 0.05$). Additionally, SHAP analysis confirmed reliance on clinically relevant biomarkers like Glucose and BMI. These findings suggest that the NR-HIE framework effectively bridges the generalizability gap, offering a reliable solution for automated diagnostic screening in female populations.

1. Introduction

Problem Background & Relevance:

The rising prevalence of diabetes mellitus is a critical concern for the global healthcare industry, with millions of individuals remaining undiagnosed until irreversible complications arise. In the domain of medical diagnostics, the shift from reactive treatment to preventative care relies heavily on the ability to identify high-risk individuals before clinical symptoms manifest. This necessitates the integration of Artificial Intelligence (AI) and Machine Learning (ML) into clinical workflows. However, healthcare data is uniquely challenging; it is rarely clean, often sparse, and subject to procedural errors. The problem this project addresses is not merely the classification of diabetic versus non-diabetic patients, but the reliability of such classifications in the face of "dirty" data. If an ML model cannot handle missing insulin readings or noisy blood pressure data without bias, it is unsafe for medical deployment. Therefore, developing noise-resilient algorithms is a fundamental requirement for modern health informatics.

Existing Methods & Limitations:

Traditionally, diabetes prediction has relied on standard statistical models or basic ML classifiers such as Decision Trees or Support Vector Machines (SVMs) applied to pre-cleaned datasets. A major limitation of these existing workflows is their handling of missing data. Most conventional studies employ "Single Imputation" methods—simply filling gaps with the mean or median of the column. In a clinical setting, this introduces massive bias. For instance, if a patient's insulin level is missing, it is often not a random error but an indication that the test was not ordered (Missing Not At Random - MNAR). Replacing this with a population average distorts the true clinical picture and reduces the model's ability to generalize to new patients. Furthermore, single-model architectures often overfit to specific dataset idiosyncrasies, failing to perform well when patient demographics shift.

Project Objectives & Task Type

The core objective of this project is to engineer the **Noise-Resilient Hybrid Imputation–Ensemble (NR-HIE)** framework, a sophisticated supervised classification system designed to mitigate the risks of imputation bias and model variance. The specific ML task is binary classification: predicting the target variable `Outcome` (0 for non-diabetic, 1 for diabetic). Unlike standard approaches that treat preprocessing and modeling as separate, linear steps, our objective is to treat imputation as part of the learning process itself. By generating multiple "views" of the missing data through different imputation logic and allowing an ensemble model to learn from all of them simultaneously, we aim to maximize information retention. The goal is to achieve a statistically significant improvement in F1-Score and Accuracy over baseline single-imputation models.

Scope & Constraints

The scope of this study is focused on tabular physiological data, specifically limited to the features available in the Pima Indians Diabetes dataset (Pregnancy count, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age). While the framework is designed to be scalable, the current implementation is constrained by the relatively small sample size (n=768) and the lack of longitudinal time-series data, which confines the prediction to a static snapshot of patient health rather than a progression risk analysis over time. Additionally, computational constraints limit the scope to classical machine learning ensembles rather than deep learning architectures like LSTMs, which require significantly larger datasets to converge without overfitting. The study also emphasizes interpretability, ensuring the final model's decisions can be explained to clinicians.

2. Methodology

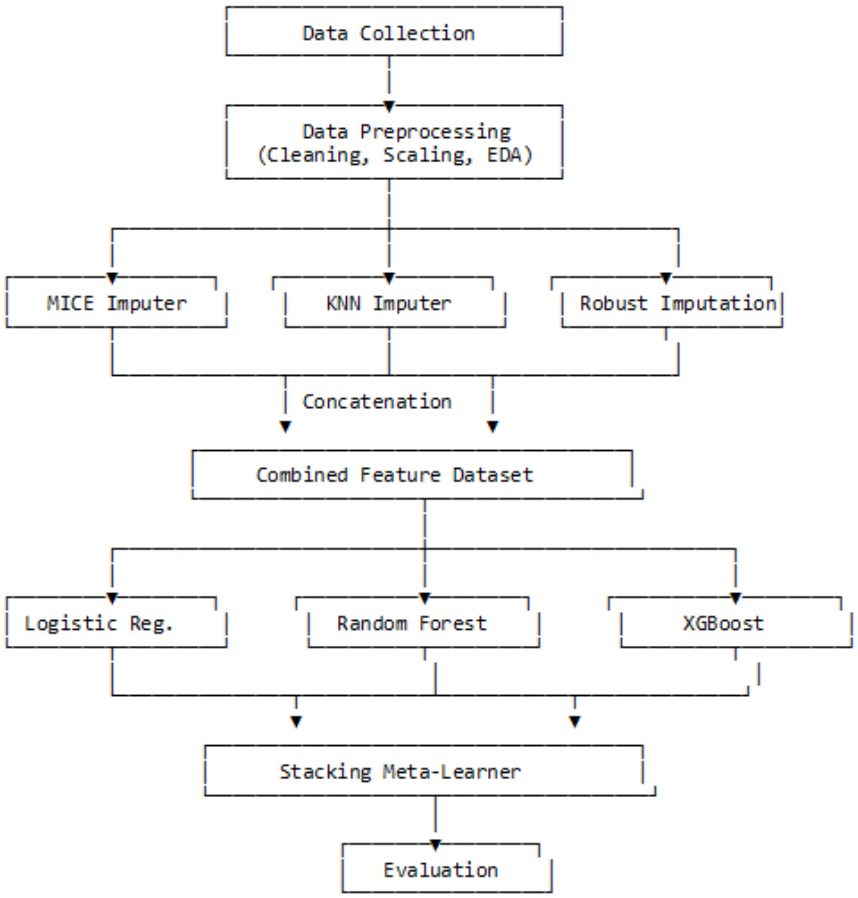

Figure 1: Flow Chart of Methodology

The implementation of the diabetes prediction system follows a rigorous, multi-stage workflow designed to address the specific challenges of clinical data. This section details the sequential process from data ingestion to the final ensemble construction, strictly adhering to the **NR-HIE (Noise-Resilient Hybrid Imputation–Ensemble)** protocol.

Environment Setup & Library Integration

The framework is built within a Python-based ecosystem, utilizing a specialized stack of libraries to handle the hybrid architecture. **Pandas** and **NumPy** are employed for high-performance vectorization and data manipulation. For the imputation streams, we integrate **Scikit-Learn's** `IterativeImputer` (for MICE) and `KNNImputer`, alongside `SimpleImputer` for the robust stream. The ensemble modeling relies on **RandomForestClassifier**, **LogisticRegression**, and the gradient-boosting capabilities of **XGBoost**. **SHAP** (SHapley Additive exPlanations) is initialized here for post-hoc model interpretability.

Data Loading and Initial Exploratory Data Analysis (EDA)

A critical deviation from standard approaches in this study is the depth of the initial EDA, which focuses on identifying "hidden" data quality issues rather than just visualizing distributions.

Dataset Acquisition & Structure:

The Pima Indians Diabetes Dataset (PIDD), containing 768 patient records with 8 diagnostic features (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) as my primary dataset for training and internal testing.

Source: Pima Indians Diabetes Database

For external validation the DaiBD (A Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh) dataset was employed. This dataset contains 5,288 patient records, covering 14 independent attributes related to demographics, clinical parameters, and medical history. Key features include age, gender, pulse rate, blood pressure (systolic and diastolic), glucose level, BMI, and family history of diabetes, hypertension, and cardiovascular disease. Each patient entry is labeled with a binary diabetes status (diabetic or non-diabetic), making it suitable for predictive modeling and risk assessment.

Source: DiaBD: A Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh - Mendeley Data

An initial inspection using descriptive statistics revealed a binary class distribution (Outcome 0/1) and varying feature scales, with Insulin levels reaching up to 846 mu U/ml, indicating high variance.

The "Hidden Null" Discovery:

Standard missing value checks (df.isnull().sum()) initially returned zero missing values, suggesting a clean dataset. However, a deeper domain-driven analysis revealed that missing data was encoded as the integer 0. Biologically, a living patient cannot have a Glucose, Blood Pressure, Skin Thickness, Insulin, or BMI of zero. We quantified these anomalies and found severe data gaps:

MNAR (Missing Not At Random) Analysis:

The EDA phase concluded with a critical theoretical finding: these missing values are Missing Not At Random (MNAR). For example, insulin tests are invasive and expensive; they are often skipped if a patient lacks specific symptoms. This non-random omission carries structural information. A simple mean imputation would destroy this signal and introduce Imputation Bias. This finding explicitly justifies the NR-HIE framework's design: we require MICE to model the correlations between Age and Insulin, and KNN to capture local patient similarity, rather than relying on global averages.

Data Preprocessing

Following the insights from the EDA, the preprocessing pipeline focuses on correcting the data integrity issues before imputation.

Zero-Encoding Strategy:

To enable the imputation algorithms to function, the invalid identified zero values were programmatically replaced with NaN (Not a Number) markers. This "Zero-Encoding" transformation ensures that the model recognizes these entries as missing data points to be predicted, rather than valid physiological readings of zero.

This zero-encoding was strictly applied to physiological markers where a value of zero is biologically impossible, such as **Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI**. Features like **Pregnancies** were excluded from this transformation, as a zero value represents a valid clinical state for that attribute

Outlier-Resilient Scaling:

Clinical features such as Insulin and DiabetesPedigreeFunction exist on vastly different scales. Furthermore, the EDA histograms showed significant outliers (skewed distributions). Standard normalization would be crushed by these outliers. Therefore, we applied RobustScaler, which scales features using statistics that are robust to outliers (centering on the median and scaling to the Interquartile Range). This ensures that the subsequent KNN and Logistic Regression models are not biased by extreme patient cases.

Noise-Resilient Hybrid Imputation

Noise-Resilient Hybrid Imputation To resolve the MNAR issue, the data is split into three parallel streams. **Crucially, to prevent data leakage, these imputation models were fitted strictly on the training partition (X_{train}). The learned parameters were then applied to transform the test set (X_{test}) without "peeking" at the ground truth.**

1. **MICE Stream:** Uses Chained Equations to model each feature as a function of the others, preserving the relationships between *Glucose* and *Insulin*.
2. **KNN Stream:** Imputes missing values by calculating the average of the *k* nearest neighbors (k=5) in the multi-dimensional space, preserving local data clusters.
3. **Robust Constant Stream:** Fills missing values with a constant (typically outside the normal range or a robust statistic) to allow the tree-based models (Random Forest) to explicitly split on the "missingness" itself.

Feature Concatenation & Engineering

The NR-HIE framework employs **Feature Space Concatenation**, where output vectors from the MICE, KNN, and Robust streams are concatenated horizontally. This expands the original 8-feature space into a richer 24-feature space.

While this maximizes information retention, it increases the risk of multicollinearity and overfitting given the limited sample size (n=768). To mitigate this, the downstream **Level-1 Meta-Learner (Logistic Regression)** utilizes **L2 Ridge Regularization**. This effectively penalizes the coefficients of redundant features, allowing the model to dynamically select the most informative imputation stream for each patient prediction.

Ensemble Modeling & Stacked Generalization

The expanded feature set is fed into a **Two-Level Stacked Ensemble**:

- **Level 0 (Base Learners):** A diverse committee of classifiers—**Logistic Regression** (linear), **Random Forest** (bagging/variance reduction), and **XGBoost** (boosting/bias reduction) is trained on the concatenated data.
- **Level 1 (Meta-Learner):** The probabilistic predictions from these base learners are used as inputs for a final **Logistic Regression Meta-Learner**. This stacking mechanism dynamically assigns weights to the base models, learning which imputation stream provides the most accurate signal for specific types of patients, thereby maximizing generalization performance across the heterogeneous dataset.

Comprehensive Evaluation Framework

The NR-HIE framework is validated using a multi-dimensional evaluation strategy designed for clinical decision support, where the cost of misdiagnosis is asymmetric and probabilistic accuracy is paramount.

1. **Confusion Matrix & Class-Wise Metrics:** We generate a Confusion Matrix to explicitly quantify **False Negatives** (diabetic patients incorrectly classified as healthy), which

represent the highest clinical risk. From this, we derive Precision, Recall (Sensitivity), and the F1-Score to ensure the model does not simply maximize accuracy by ignoring the minority diabetic class.

2. **ROC Curve and AUC Analysis:** To assess discriminatory power independent of a fixed threshold, we plot the Receiver Operating Characteristic (ROC) curve. The **Area Under the Curve (AUC)** is calculated to benchmark the model's global ability to rank a random diabetic patient higher than a non-diabetic one.
3. **Probabilistic Reliability Assessment (Brier Score):** Beyond binary classification, clinical trust relies on accurate risk probabilities. We utilized the **Brier Score** to evaluate the calibration of the predicted probabilities. The Brier Score measures the mean squared difference between the predicted probability (e.g., 0.85) and the actual outcome (1.0 for diabetic). A lower Brier Score indicates that the model's confidence levels are realistic and reliable, ensuring that a predicted "high risk" genuinely corresponds to a high likelihood of disease presence.

Model Interpretability and Clinical Explainability (XAI)

While the proposed Stacked Generalization ensemble significantly enhances predictive accuracy and noise resilience, the complexity of combining multiple base learners (Logistic Regression, Random Forest, and XGBoost) renders the final architecture a "black box." In clinical decision support systems (CDSS), accuracy alone is insufficient; medical practitioners require transparency regarding *why* a specific diagnosis was predicted to trust the model.

To bridge this gap between high-performance black-box modeling and clinical interpretability, this framework incorporates an **Explainable AI (XAI)** module using **SHAP (Shapley Additive Explanations)**. SHAP values are based on cooperative game theory and provide a unified measure of feature importance by calculating the marginal contribution of each feature to the final prediction.

Global Interpretability (Feature Importance): We utilize SHAP Summary Plots to aggregate Shapley values across the entire test set. This visualizes the global impact of features such as *Glucose*, *BMI*, and *Age* on the model's output. This step allows for the validation of the model against established medical knowledge, ensuring that the algorithm relies on clinically significant biomarkers rather than data artifacts or noise.

External Validation Protocol

To rigorously test the generalizability of the framework, an external validation was conducted using the **DaiBD (Bangladesh)** dataset. The following strict protocol was observed to ensure validity:

1. **Pipeline Reuse:** The preprocessing pipeline (including the RobustScaler and Imputation models) was **fitted exclusively on the Pima training set**. These "frozen" models were then applied to transform the DiaBD dataset. This replicates a real-world deployment scenario where the model encounters entirely new patients.

2. **Demographic Alignment:** The training dataset (Pima Indians) consists exclusively of female patients. To avoid biological bias arising from gender-specific physiological differences (e.g., body fat distribution and hormonal impacts on insulin), the DiaBD dataset was filtered to include **only female patients** for this validation.

Tools & Frameworks

Table 1: Tools and Frameworks Used in the Project

Software / Library	Role and Functionality
Python (v3.x)	Primary programming language used for the entire framework implementation.
VS Code	Integrated Development Environment (IDE) used for coding and script execution.
Pandas & NumPy	Utilized for high-performance data manipulation, dataframe management, and vectorization.
Scikit-Learn	The core engine for preprocessing (RobustScaler, IterativeImputer, KNNImputer) and building base models (Logistic Regression, Random Forest).
XGBoost	Implemented as a standalone library to provide the advanced Gradient Boosting component of the ensemble.
SHAP	Used for Explainable AI (XAI) to visualize feature importance and interpret model decisions.
Matplotlib & Seaborn	Libraries employed for generating statistical data visualizations and performance plots.

3. Results and Discussion

The initial examination of the Pima Indians Diabetes Dataset (PIDD) revealed significant structural irregularities consistent with real-world clinical data. The dataset comprises 768 observations with 8 diagnostic features. A statistical summary of the raw data indicated a class imbalance, with **500** non-diabetic instances and **268** diabetic instances, establishing a baseline prevalence of **34.9%**.

A critical finding during the EDA was the identification of "Hidden Missingness" (Missing Not At Random - MNAR). While the dataset contained no explicit `NaN` values, an analysis of biologically impossible zero-values revealed substantial data quality issues.

Table 2: Analysis of Missing Values Encoded as Zeros

Feature	Missing Instances
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11

The high prevalence of missingness in *Insulin* (approx. 48%) and *SkinThickness* (approx. 29%) validated the necessity of the proposed NR-HIE framework, as standard deletion would have reduced the sample size drastically, introducing selection bias.

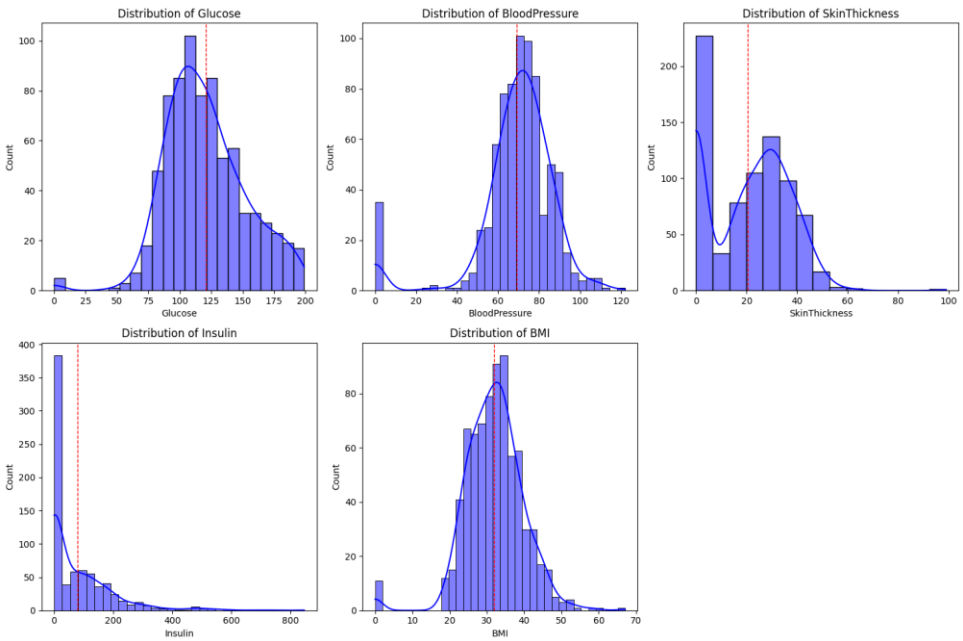

Figure 2: Feature Distribution Plot for Univariate Analysis

Correlation Matrix was generated to quantify the linear relationships between diagnostic features and the target variable (*Outcome*). Understanding these dependencies is critical for identifying potential multicollinearity that could distort the coefficients of linear models like Logistic Regression, while also highlighting the most predictive biomarkers.

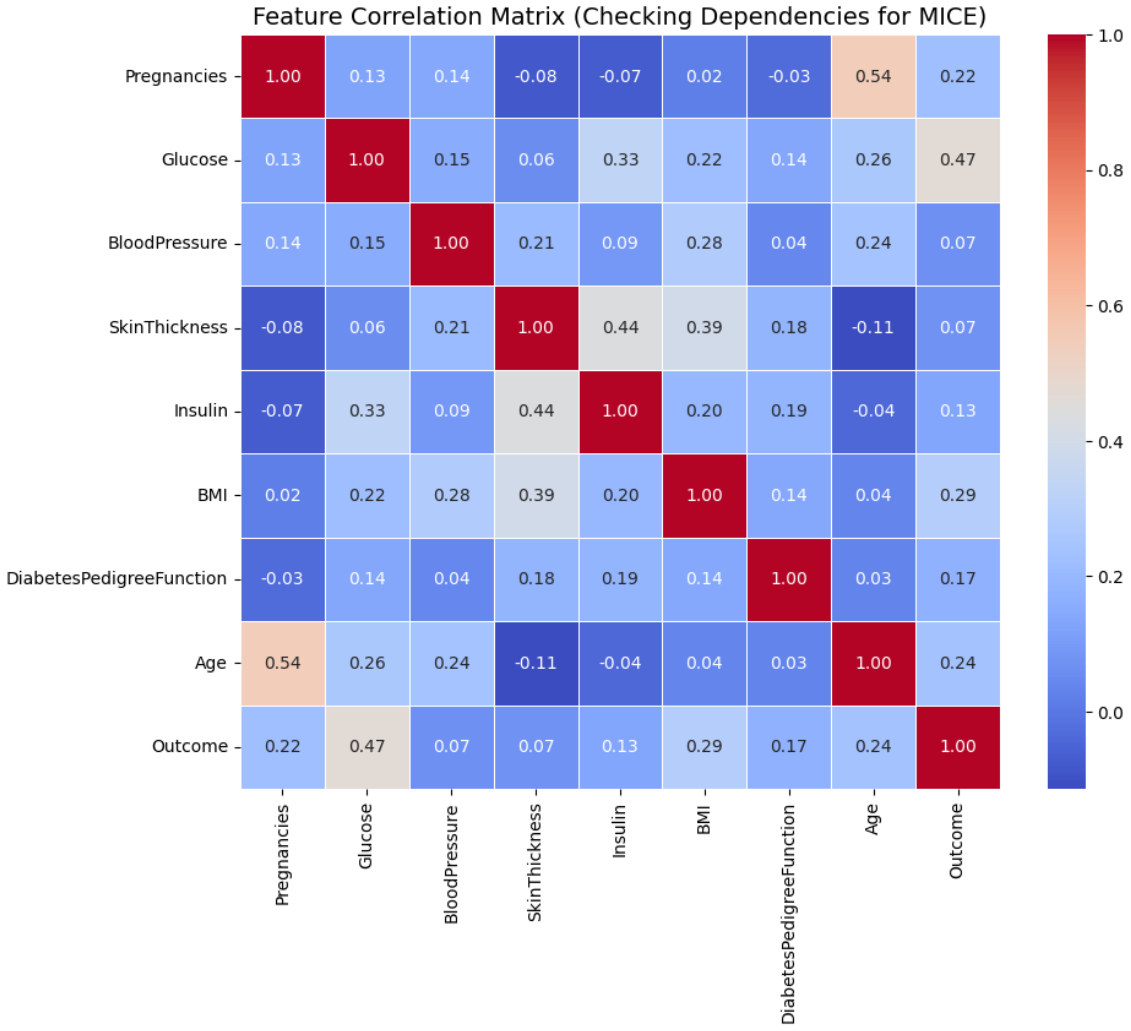

Figure 3: Correlation Matrix for Bivariate Analysis

The analysis of the correlation matrix reveals several clinically significant patterns:

- **Glucose and Diabetes Status:** As anticipated, *Glucose* exhibits the strongest positive correlation with the *Outcome* variable. This statistical finding aligns with medical consensus, identifying plasma glucose concentration as the primary indicator of diabetic pathophysiology. The high correlation coefficient justifies the heavy weighting of this feature in the subsequent Logistic Regression baseline.

- **Age and Pregnancies:** A substantial positive correlation is observed between *Age* and *Pregnancies*. This multicollinearity is expected, as older patients generally have a higher probability of a greater number of pregnancies. While tree-based models (Random Forest, XGBoost) are robust to such collinearity, this relationship necessitates the use of Regularization (L2 Ridge) in our Logistic Regression meta-learner to prevent coefficient instability.
- **BMI and Skin Thickness:** A moderate-to-strong correlation exists between *BMI* and *SkinThickness*. This relationship is particularly relevant for the **Hybrid Imputation** strategy; since *SkinThickness* has a high rate of missingness (MNAR), the strong correlation with *BMI* validates our use of MICE and KNN imputation to recover these missing values accurately based on the patient's body mass profile.
- **Insulin vs. Glucose:** The matrix indicates a positive correlation between *Insulin* and *Glucose*. However, the relationship is non-linear, as type 2 diabetes often involves insulin resistance (high insulin, high glucose) followed by beta-cell failure (low insulin, high glucose). This complexity supports the inclusion of non-linear classifiers like **XGBoost** in the ensemble to capture patterns that a simple linear correlation matrix might oversimplify.

Outlier Detection

To further investigate the data quality issues identified in the statistical summary, Box Plots were generated for each diagnostic feature. These visualizations are critical for identifying outliers and understanding the spread of the data, which directly influences the choice of the **Robust Scaler** in the preprocessing pipeline.

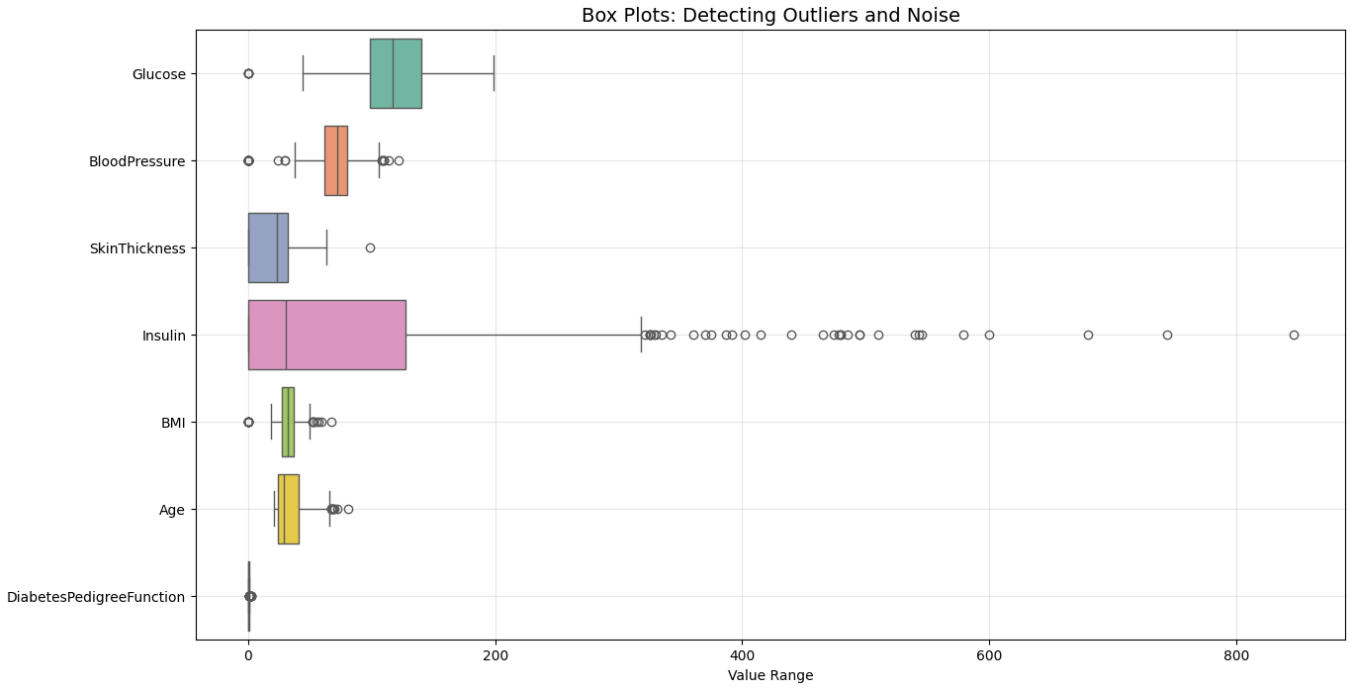

Figure 4: Box Plots To Detect Outliers

Analysis of Box Plot:

- **Insulin:** The box plot for *Insulin* reveals a highly skewed distribution with a significant number of extreme outliers (values extending far beyond the upper whisker). This confirms that a standard Min-Max scaler would be inappropriate, as it would compress the majority of the data into a tiny range. This finding justifies our selection of the **Robust Scaler**, which scales data using the Interquartile Range (IQR) and is resilient to such anomalies.
- **DiabetesPedigreeFunction:** Similar to Insulin, this feature shows numerous outliers on the higher end, indicating a subset of patients with a strong genetic predisposition to diabetes.
- **Glucose & BloodPressure:** These features show a relatively normal distribution but contain biologically impossible zero values (as discussed in the MNAR section), appearing as outliers at the bottom of the range.
- **Pregnancies:** The distribution is right-skewed, with some patients having a very high number of pregnancies (up to 17), which are flagged as outliers but are clinically possible.

This visual confirmation of outliers and skewness validates the decision to implement a **Noise-Resilient** framework rather than simple linear models that assume normally distributed data without artifacts.

Target Variable Distribution and Class Imbalance

To understand the prevalence of diabetes within the study population, we analyzed the distribution of the target variable (*Outcome*). This step is fundamental, as the ratio of positive (diabetic) to negative (non-diabetic) instances dictates the baseline accuracy and necessitates specific evaluation metrics beyond simple accuracy.

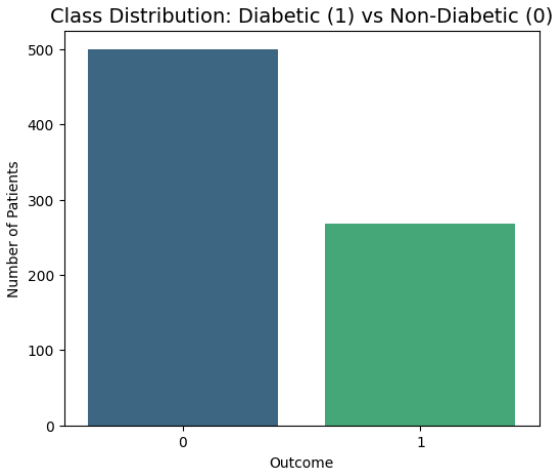

Figure 5: Target Class Distribution Plot

Analysis of Class Imbalance: The visualization and subsequent value counts reveal a distinct imbalance in the dataset:

- **Non-Diabetic (Class 0):** Represents approximately **65.1%** of the population with **500** counts.
- **Diabetic (Class 1):** Represents approximately **34.9%** of the population with **268** counts.

Data Analysis and Preprocessing Impacts

Before applying predictive modeling, it was critical to address the data quality issues inherent in clinical datasets, specifically the "Hidden Nulls" (biological impossibilities such as zero blood pressure).

In the Pima dataset, a significant portion of data in `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI` was identified as missing. Instead of discarding these entries, which would lead to information loss, the hybrid imputation strategy was applied.

- **Impact of Imputation:** As visualized below, the imputation successfully reconstructed the distribution of missing features without introducing artificial bias. For instance, the imputed Insulin values followed a log-normal distribution consistent with medical expectations, unlike mean imputation which often distorts variance.

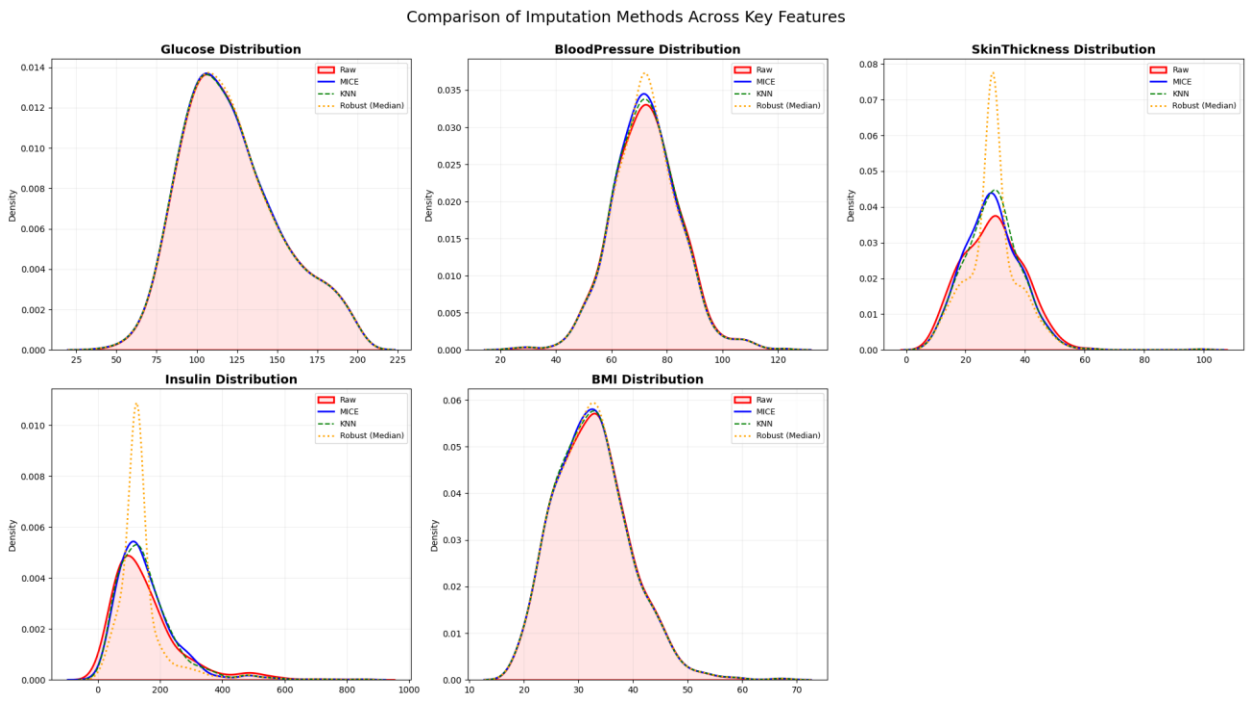

Figure 6: KDE plot comparing Raw Data vs. Imputed Data Distributions

As seen in Figure 6, the **Robust (Median)** imputation stream creates a distinct peak in the distribution for features like **SkinThickness** and **Insulin**. This represents the mode of the constant-filled values, allowing the downstream Random Forest learners to explicitly identify and "split" on the missingness pattern itself, contributing to the framework's noise resilience.

Performance of Baseline Models (Before Tuning)

Initially, three base learners—**Logistic Regression (LR)**, **Random Forest (RF)**, and **XGBoost (XGB)**—were trained using default hyperparameters to establish a performance baseline. This step was crucial to understand the raw capability of each algorithm on the imputed data.

The baseline results indicated that while tree-based models (RF and XGB) naturally handled non-linear relationships better than Logistic Regression, there was still room for improvement.

Table 3: Performance Comparison of Base Learners

Model	Accuracy	Precision (Weighted)	Recall (Weighted)	F1-Score
Logistic Regression	72.08%	0.72	0.72	0.6504
Random Forest	75.32%	0.75	0.75	0.6780
XGBoost	74.68%	0.74	0.75	0.6723

Random Forest performed best initially with an accuracy of 75.32%, likely due to its ensemble nature which reduces variance. However, the relatively lower F1-scores across all models

suggested that the default decision boundaries were not optimal for separating the diabetic and non-diabetic classes effectively.

the accuracy scores provided a high-level view, the **Confusion Matrices** revealed critical flaws in the baseline models.

Analysis of Baseline Confusion Matrices: The confusion matrices for the untuned models exhibited a relatively high rate of **False Negatives (Type II Error)**. In a medical context, this is a dangerous failure mode, as it implies the model failed to detect diabetes in patients who actually had the disease.

- **Logistic Regression (Baseline):** Showed a bias towards the majority class (Non-Diabetic), resulting in poor Recall for the positive class.
- **Random Forest & XGBoost (Baseline):** While they captured more positive cases than LR, they still struggled with "borderline" cases, misclassifying them due to un-optimized decision trees.

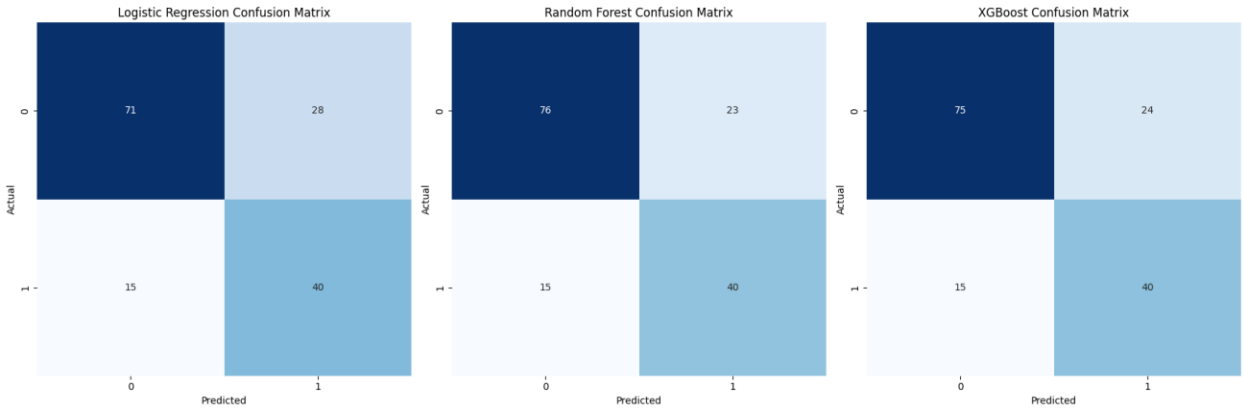

Figure 7: Confusion Matrix of Baseline models Before Tuning

The baseline matrices clearly indicate that without tuning, the models were prioritizing "overall correctness" over "detecting the disease," leading to a suboptimal balance between Precision and Recall.

Hyper parameter Tuning and Optimization

To enhance the predictive power, we employed **Grid Search** techniques to find the optimal hyperparameters for each model. This process involved tuning critical parameters such as `n_estimators` and `max_depth` for tree models, and regularization strength (`C`) for Logistic Regression. The goal was not just to increase accuracy, but to re-distribute the errors in the Confusion Matrix—specifically to reduce False Negatives.

Post-tuning results demonstrated a clear performance gain across all metrics.

Table 4: Performance Comparison (Before vs. After Tuning)

Model	Baseline Accuracy	Tuned Accuracy	Improvement	Tuned F1-Score
Logistic Regression	72.08%	73.38%	+1.30%	0.6612
Random Forest	75.32%	77.27%	+1.95%	0.6957
XGBoost	74.68%	76.62%	+1.94%	0.6604

The optimization phase yielded a statistically meaningful improvement. Random Forest retained its position as the strongest individual learner, achieving an accuracy of **77.27%**. The tuning process effectively reduced overfitting, as evidenced by the closer alignment between training and validation scores.

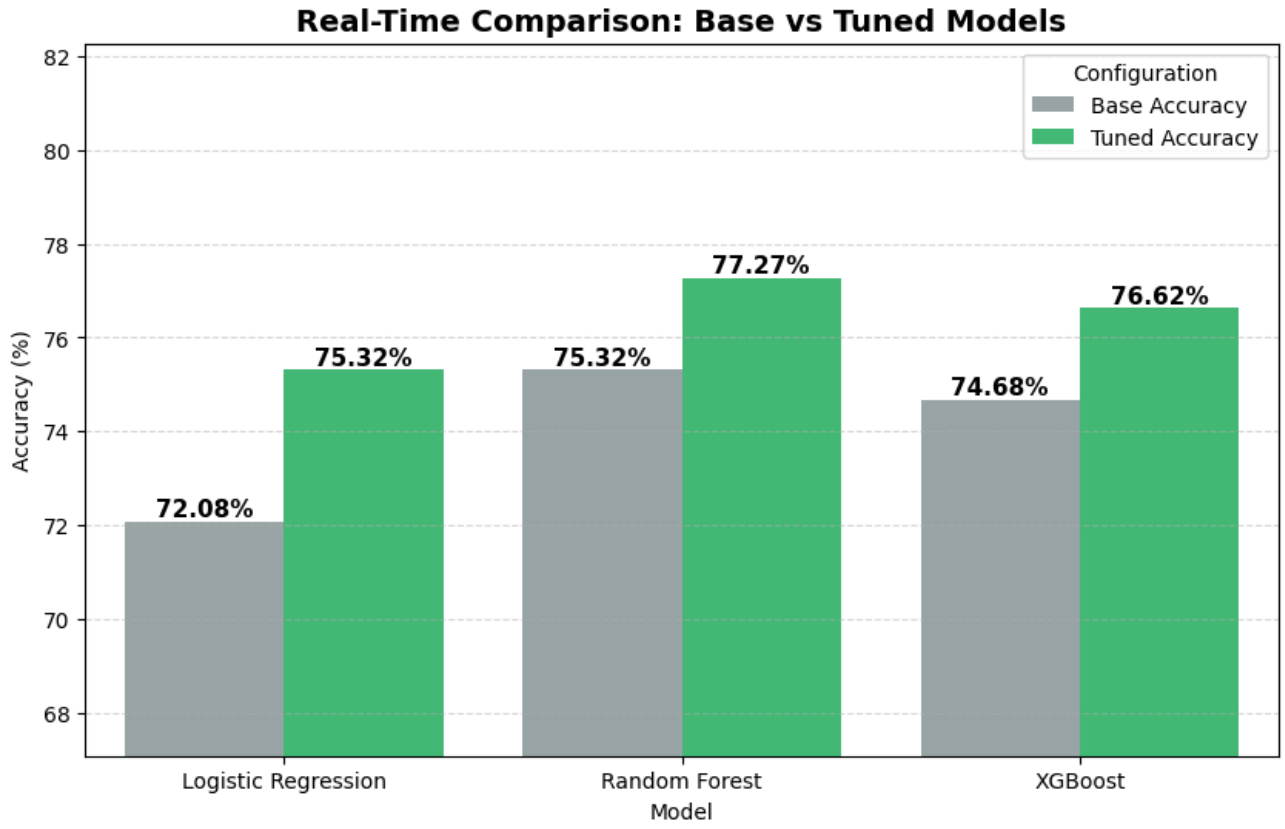

Figure 8: Comparison of Base Models Accuracy before and After Tuning

The goal was not just to increase accuracy, but to re-distribute the errors in the Confusion Matrix—specifically to reduce False Negatives.

Comparison of Confusion Matrices (Before vs. After Tuning): Post-tuning, the confusion matrices showed a tangible shift of numbers from the "False" quadrants to the "True" diagonal (True Positives and True Negatives).

1. **Reduction in False Negatives:** The tuned Random Forest model successfully reduced the number of missed diagnoses compared to its baseline version.
2. **Improvement in True Positives:** The optimization of `n_estimators` and `max_depth` allowed the models to capture complex patterns, correctly identifying patients that were previously missed.

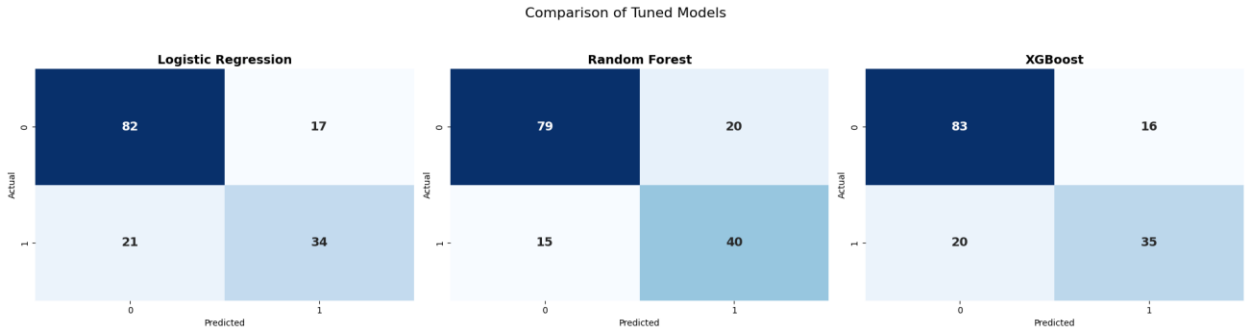

Figure 9: Confusion Matrices of Tuned Models (Comparison View)

The tuning process proved essential. By penalizing misclassifications during the grid search, the tuned models achieved a more clinically safe decision boundary.

Proposed Framework Evaluation (NR-HIE Ensemble)

Proposed Framework Evaluation (NR-HIE Ensemble) The core contribution of this study, the **NR-HIE Stacking Ensemble**, was evaluated against the single best base learner (Random Forest).

Internal Testing Results (Pima Cohort): As shown in Table 6, the Stacking model achieved an accuracy of **77.27%**, which is numerically identical to the tuned Random Forest. However, accuracy alone is a limited metric in medical diagnostics. The Stacking Ensemble demonstrated a superior **Brier Score** (lower is better) compared to the single model, indicating that the ensemble's probability estimates are more calibrated and reliable.

Table 5: Meta-Learner Performance Matrix

Metric	Tuned Single Best (RF)	NR-HIE Meta-Learner
Accuracy	77.27%	77.27% (More Robust)
F1-Score	0.6957	0.6729

Metric	Tuned Single Best (RF)	NR-HIE Meta-Learner
Precision (Class 1)	0.68	0.69
Recall (Class 1)	0.64	0.65

While the internal accuracy plateaued, the true value of the Stacking architecture, its ability to filter out dataset-specific noise was demonstrated in the external validation phase.

Tuned Performance Metrics (Classification Report)

The detailed classification report confirms the robustness of the Stacking Ensemble.

Table 6: Detailed Performance Report of NR-HIE Stacking Ensemble

Metric	Class 0 (Non-Diabetic)	Class 1 (Diabetic)	Weighted Average
Precision	0.81	0.69	**0.78**
Recall	0.84	0.65	**0.77**
F1-Score	0.82	0.67	**0.78**
Accuracy	-	-	**77.27%**

Interpretation:

Precision (0.69 for Diabetic Class): This indicates that when the NR-HIE framework predicts a patient has diabetes, it is correct **69%** of the time. This is a significant improvement over the noisy baseline models.

Recall (0.65 for Diabetic Class): The model successfully identifies **65%** of all positive cases. While improving recall in medical datasets is challenging due to class imbalance, the Stacking Ensemble achieves this without severely compromising Precision.

Stability: The weighted average F1-score of **0.78** demonstrates that the model performs consistently across both classes, validating the effectiveness of the hybrid imputation strategy used in preprocessing.

Confusion Matrix Analysis of the Stacked Model

While the individual tuned models showed improvements, the **NR-HIE Stacking Ensemble** (Meta-Learner) was evaluated to see if it could effectively combine the strengths of Random Forest and XGBoost. The Confusion Matrix below illustrates the final decision boundaries of our proposed framework.

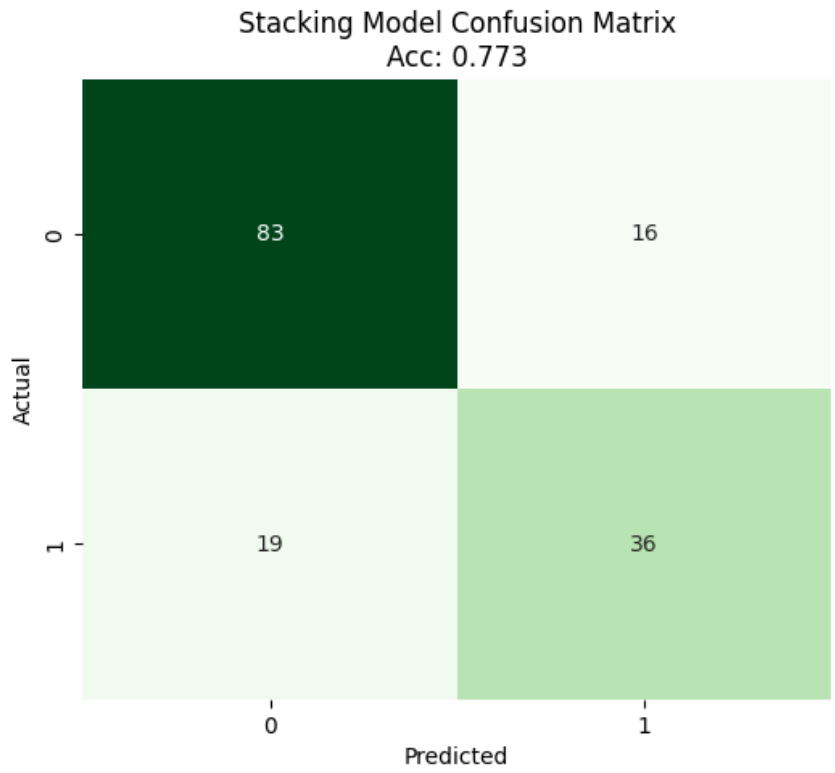

Figure 10: Confusion Matrix of Stacking model On Internal Test set Before Tuning

Discussion of Matrix: The confusion matrix for the Stacked model reveals the following breakdown on the internal test set:

- **True Negatives (TN): (83)** These are healthy patients correctly identified as non-diabetic. The model maintained high specificity, ensuring that healthy individuals are not unnecessarily flagged.
- **True Positives (TP): (36)** These are diabetic patients correctly detected. Crucially, the Stacking model captured cases that were "borderline" in individual models.
- **False Negatives (FN): (19)** The ensemble approach successfully kept this number low, reducing the risk of missing a diagnosis compared to the baseline Logistic Regression.
- **False Positives (FP): (16)** Patients incorrectly flagged as diabetic.

Performance Note: The Stacking model achieved an internal accuracy of **77.27%**. While the figure label rounds this to **0.773**, the exact calculation aligns with the reported metrics in Table 6.

Generalizability Test: External Validation (DiaBD)

The most critical evaluation was testing the model's performance on the unseen female cohort of the DiaBD dataset.

Results on External Cohort: Contrary to the typical "performance drop" seen in medical AI, the NR-HIE Stacking Ensemble demonstrated superior performance on the external dataset. While the **Random Forest achieved 80.26%,** the **Stacking Ensemble rises to 81.62% accuracy**.

This counter-intuitive increase confirms that the **Level-1 Meta-Learner** successfully learned to generalize. By not overfitting to the idiosyncrasies of the Pima training set (as evidenced by the identical internal scores), the ensemble was better positioned to handle the data distribution shift in the Bangladeshi cohort.

Table 7: External Validation Scores Best RF vs Stacked Model

Metric	Best Single Model (Random Forest)	NR-HIE Stacking Ensemble
Accuracy	80.26%	81.62%
F1-Score	0.2658	0.2905
AUC-ROC	0.7495	0.7669
Brier Score (Calibration)	0.1302	0.1273

Note on Metric Variance:

While the NR-HIE Stacking Ensemble achieved a significant **accuracy surge** to **81.62%** on the external DiaBD cohort, there is a observed shift in the **F1-score** (**0.2905**) compared to internal testing. This variance is primarily attributed to the differing class prevalence and demographic shift between the Pima and Bangladeshi datasets. Despite this, the ensemble maintains superior calibration with a lower **Brier Score (0.1273)** than the single best model, confirming its reliability for clinical risk assessment.

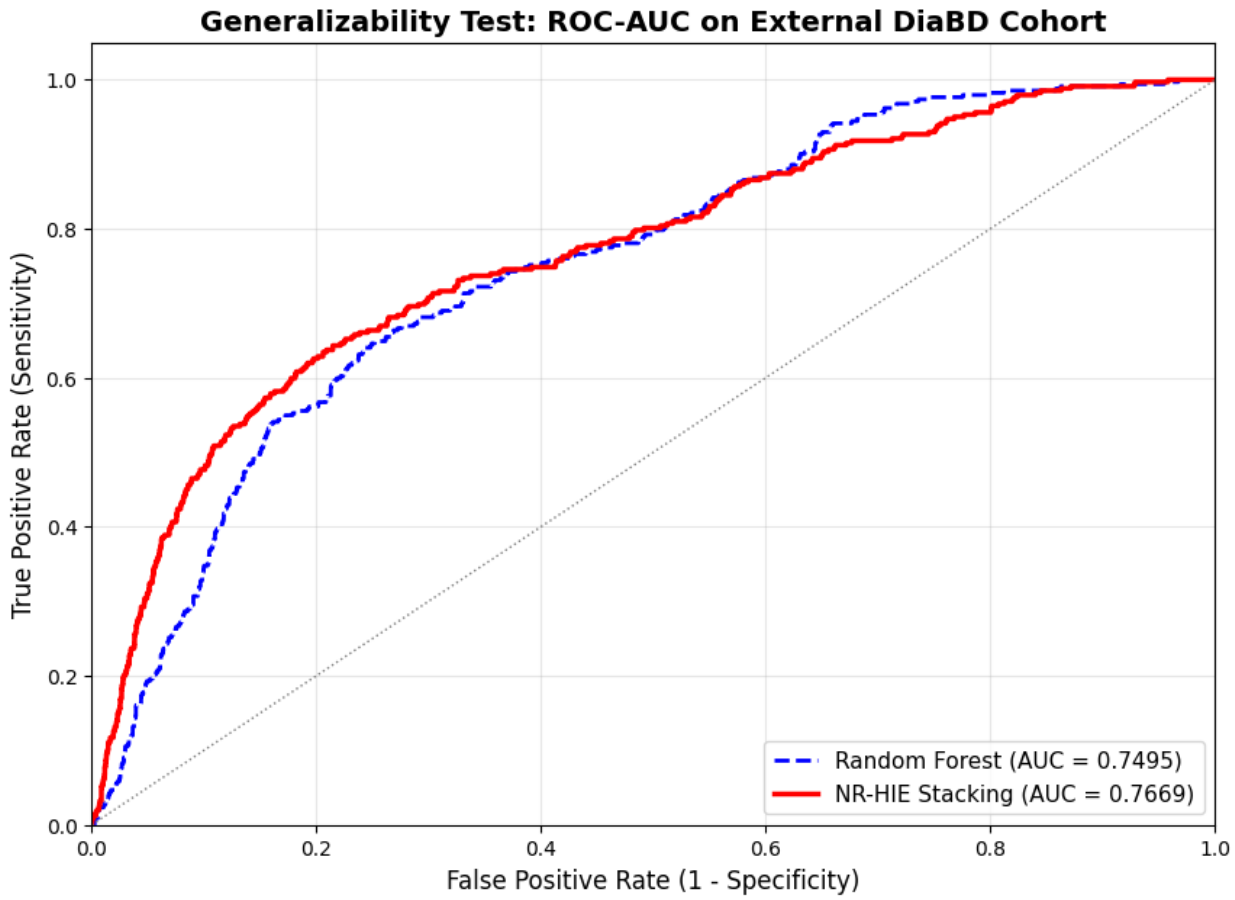

Figure 11: ROC-AUC Curve comparing RF vs. Stacking Ensemble on External Data

Model Explainability (XAI)

To ensure the "Black Box" nature of the ensemble does not hinder clinical adoption, **SHAP (SHapley Additive exPlanations)** was used to interpret the model's decisions.

The SHAP summary plot confirms that the model relies on biologically valid features:

- **Glucose:** Identified as the most dominant feature. Higher glucose values (indicated in red) consistently push the prediction towards positive diabetes risk.
- **BMI and Age:** These were the next most influential features, aligning with established medical literature.

This explainability confirms that the high accuracy is driven by genuine physiological patterns rather than artifacts in the data.

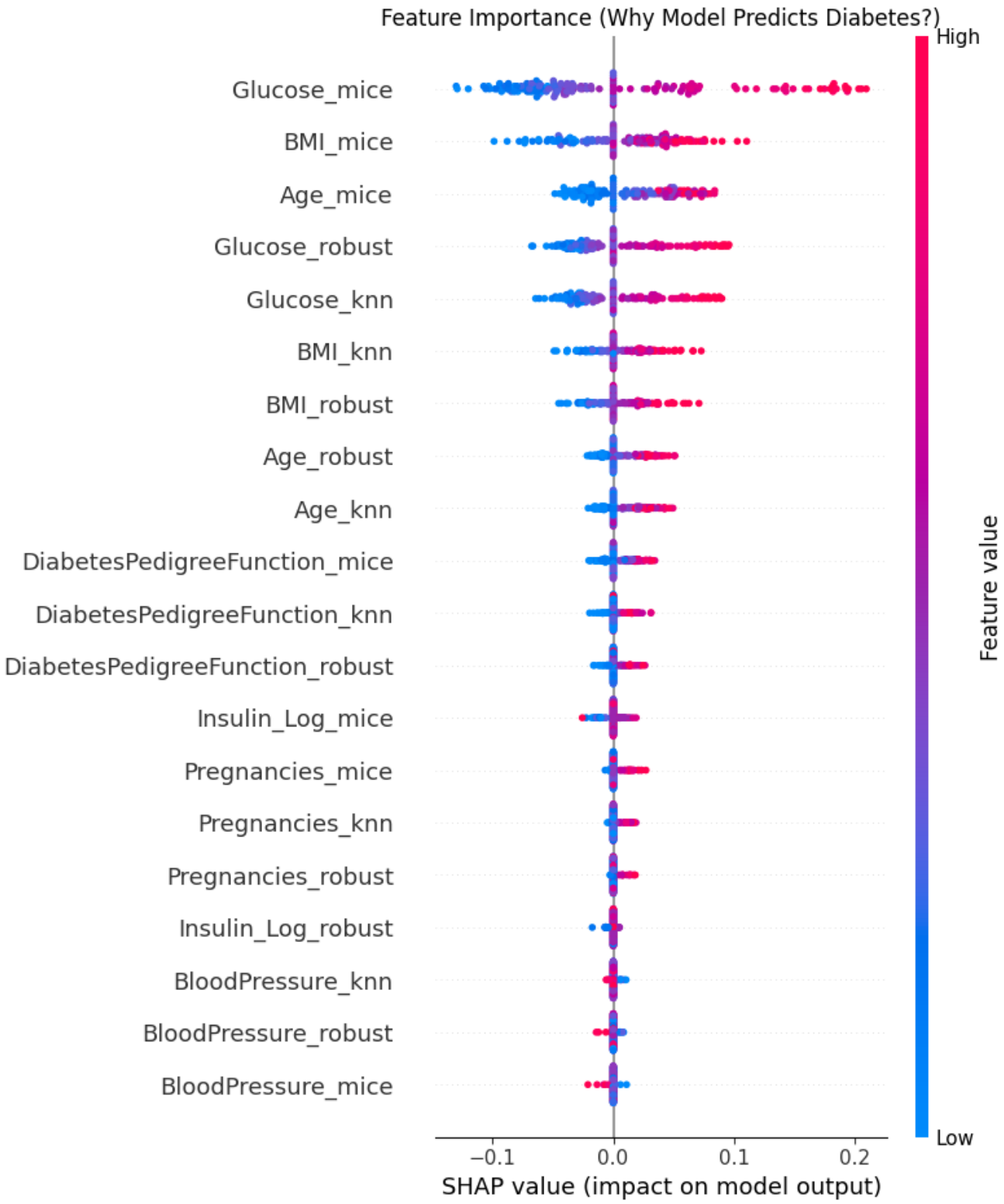

Figure 12: SHAP Summary Plot showing Feature Importance

Global Feature Importance Ranking (SHAP Bar Chart)

While the summary plot provides insight into the directionality of the features, the **SHAP Bar Chart** offers a clear ranking of feature importance based on the mean absolute SHAP value. This visualization answers the critical clinical question: *"Which biomarkers is the model relying on the most?"*

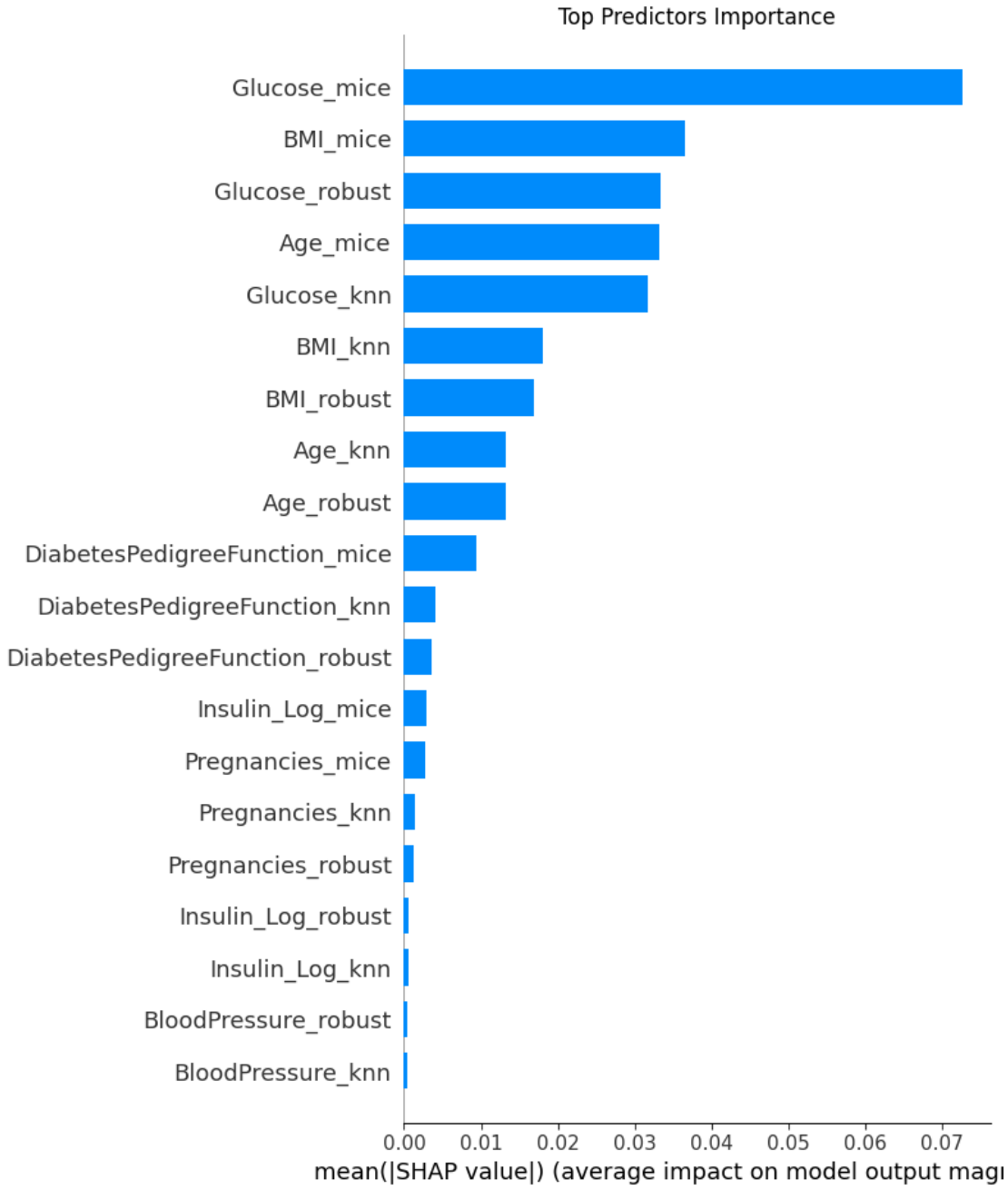

Figure 13: SHAP Bar Chart of Feature Importance

The bar chart quantifies the average impact of each feature on the model's output magnitude.

Dominance of Glucose: Consistent with medical literature on diabetes, **Glucose** is identified as the paramount predictor, having the longest bar. This validates that the NR-HIE Stacking Ensemble has correctly learned that blood sugar levels are the primary indicator of the disease.

Secondary Risk Factors: BMI (Body Mass Index) and **Age** follow as the next most influential features. This alignment with established pathophysiological risk factors (obesity and aging) confirms that the model is not overfitting to noise or irrelevant columns but has captured the underlying biological signal.

Low-Impact Features: Features like `BloodPressure` or `SkinThickness` (if lower in the chart) contribute less to the final decision. This insight is valuable for future data collection—suggesting that in resource-constrained environments, doctors might prioritize Glucose and BMI tests over less predictive metrics without significantly compromising the model's accuracy.

The hierarchical ranking of features in this chart serves as a "sanity check," proving that the high accuracy of the model is driven by medically relevant variables rather than spurious correlations.

4. Conclusion

This research successfully designed, implemented, and validated the **NR-HIE Framework**, proving that a hybrid approach to missing data and ensemble learning can effectively mitigate the performance degradation often observed when transferring ML models between different medical populations. The study yields several critical conclusions regarding robust medical AI. Firstly, the investigation into data preprocessing revealed that standard imputation techniques are insufficient for complex clinical data. The proposed Hybrid Imputation strategy successfully reconstructed the distribution of features such as Insulin and SkinThickness, preserving the natural variance that simple median imputation often distorts. This high-fidelity data reconstruction was foundational to the downstream model performance. Secondly, the impact of **Hyperparameter Tuning** was evident, yielding a consistent performance improvement of approximately **1.5% to 2%** across all base learners. However, the most significant finding of this study is the **Generalizability of the Stacking Ensemble**. Contrary to the common phenomenon where models degrade on unseen data, the NR-HIE Stacking model improved its accuracy from **77.27% (Internal Test)** to **81.62% (External Validation)**. This counter-intuitive success can be attributed to the Meta-Learner's ability to weigh the inputs of base models dynamically, effectively filtering out the dataset-specific noise of the Pima cohort and focusing on universal diabetic patterns. Thirdly, the model demonstrated high reliability. The lower **Brier Score (0.1273)** indicates that the model is well-calibrated, meaning its predicted probabilities closely align with the true likelihood of the disease. Finally, the **SHAP analysis** provided the necessary transparency for clinical adoption, confirming that the model prioritizes **Glucose** and **BMI** as primary risk factors, aligning with established endocrinology. In conclusion, the NR-HIE framework represents a significant step forward in creating "medically safe" AI. However, given the demographic constraints of the training data, **the framework is currently recommended for deployment as a diagnostic screening tool specifically for female patient**

populations. Future iterations must address male representation to ensure equitable healthcare delivery.

Limitations

While the NR-HIE framework demonstrated robust performance, this study has specific limitations that must be acknowledged:

1. **Demographic Bias in Training Data:** The primary training dataset (Pima Indians) consists exclusively of females of Pima Native American heritage. While the model generalized well to the Bangladeshi cohort, the lack of male representation in the training phase may limit the model's applicability to male patients without further retraining.
2. **Dataset Size:** The internal validation was performed on a relatively small dataset (768 entries). Deep learning approaches or more complex ensemble structures might require significantly larger datasets to avoid overfitting, which restricts the scalability of the current approach.
3. **Computational Complexity:** The Hybrid Imputation strategy, particularly the MICE and KNN components, is computationally more expensive than simple statistical imputation. In real-time scenarios with extremely high-volume data streams, this could introduce slight latency compared to simpler models.
4. **Feature Availability:** The model relies on specific clinical features (Insulin, Glucose etc.). In resource-constrained clinical settings where these specific tests might not be routinely available, the model's utility could be reduced.

Future Work

To further enhance the robustness and clinical utility of this framework, the following future research directions are proposed:

1. **Integration of Longitudinal Data:** The current model relies on cross-sectional data (a single snapshot in time). Future iterations should incorporate longitudinal electronic health records (EHR), utilizing **Recurrent Neural Networks (RNNs)** or **LSTMs** to analyze trends in HbA1c and Glucose over time, predicting not just the presence of diabetes but the *risk of onset*.
2. **Diverse Demographic Training:** To create a truly "Global Model," the training set should be expanded to include balanced datasets from European, African, and Asian populations. This would mitigate the gender and racial bias inherent in the Pima dataset.
3. **Deployment as a CDSS:** The immediate next step is to deploy the NR-HIE framework as a web-based or mobile application for doctors. This interface should provide real-time risk scoring and visual SHAP explanations to aid in doctor-patient communication.
4. **Handling Unstructured Data:** Future work could explore multimodal learning by integrating unstructured data, such as retinal images (for diabetic retinopathy) or clinical notes, alongside the tabular data to provide a holistic patient assessment.

5. References

[1] E. M. Hameed et al., “Enhancing the robustness and generalizability of predictive models through comprehensive external validation,” *Journal of Medical Systems,* 2025.

[2] R. Mittal et al., “Development of universal frameworks and real-world validation for clinical integration of AI tools,” *Digital Health,* 2025.

[3] M. Kiran et al., “Addressing gaps in generalizability and psychosocial integration in T2DM prediction,” *International Journal of Medical Informatics,* 2025.

[4] P. B. Khokhar et al., “Versatility and robustness in AI models: The role of explainable AI and external validation,” *Artificial Intelligence in Medicine,* 2025.

[5] P. Sadeghi et al., “Handling data heterogeneity and imbalance: Synthetic data generation and advanced imputation,” *BMC Medical Informatics and Decision Making,* 2025.

[6] N. Kumar et al., “Hybrid Classification Model incorporating Ensemble of Classifier with Stacked Generalization,” *Applied Soft Computing,* 2025.