

Bridging the Generalizability in Diabetes Prediction: A Noise-Resilient-Hybrid-Imputation-Ensemble (NR-HIE) Framework for Robust Clinical Decision Support



Silicone Global Tech Gilgit Baltistan

Hasnain Alam

alamhasnain457@gmail.com

+92-355-4286542

Department: Artificial Intelligence and Data Science

Supervisor: Hafizuddin

Abstract

Machine learning models designed for diabetes prediction often demonstrate high performance on controlled academic datasets but experience a significant drop in efficiency when deployed in real-world clinical settings-a phenomenon widely known as the Generalizability Gap. This decline is fundamentally driven by factors such as data heterogeneity, non-random missingness (MNAR), and the imputation bias introduced by single imputation methods. This proposal introduces the Noise Resilient Hybrid Imputation–Ensemble (NR-HIE) Framework, specifically engineered to address and mitigate this core imputation bias. The NR-HIE framework employs a parallel architecture featuring three distinct imputation streams: Multiple Imputation by Chained Equations (MICE) for complex dependencies, K-Nearest Neighbors (KNN) for local structure-aware filling and Robust/Constant Imputation for noise resilience. The outputs from these streams are integrated via feature-space concatenation, which serves as input to a diversified ensemble of base learners, including Logistic Regression, Random Forest, and XGBoost. The final prediction synthesis utilizes stacked generalization, a meta-learning strategy proven to enhance cross-dataset stability. Crucially, the system incorporates comprehensive external validation and Explainable AI (XAI) using SHAP value analysis to ensure not only predictive accuracy but also transparency and clinical interpretability. By combining noise-resilient data handling with advanced ensemble techniques and rigorous validation, the NR-HIE framework aims to deliver a robust, stable, and clinically actionable diabetes prediction model, capable of maintaining consistent performance across diverse populations and healthcare infrastructures.

1. Introduction

Machine learning models have shown promise in predicting diseases like diabetes, frequently achieving high reported accuracies on controlled benchmark datasets. A major impediment to their adoption, however, is the Generalizability Gap—the inability of these models to sustain high performance when translated into real-world clinical practice.

Clinical datasets are inherently messy, characterized by multi-source biases, significant heterogeneity, and crucially, non-random missingness (Missing Not at Random or MNAR) patterns. Traditional data preprocessing often relies on simplistic imputation methods that assume data is Missing at Random (MAR), inadvertently introducing a critical imputation bias that severely compromises the downstream model's stability and portability across different patient cohorts and institutions.

The evolution of clinical Artificial Intelligence (AI) demands a shift from benchmark-oriented experimentation toward the creation of robust, universal frameworks that genuinely support real-world decision-making. Recent evaluations of medical AI emphasize that models optimized under controlled conditions seldom maintain efficacy across varying populations, geographical boundaries, and diverse healthcare infrastructures. Therefore, ensuring prediction robustness requires the implementation of comprehensive external validation strategies.

For diabetes prediction, where early, accurate detection is vital for prognosis, methodological innovations must be introduced to withstand data noise, class imbalance, and structural variation inherent in clinical data. This proposal introduces the critical challenge of Imputation Bias, which arises from inconsistent and assumption-dependent handling of heterogeneous clinical data.

To this end, the project establishes four core objectives: (1) To Implement Comprehensive External Validation consistent with recommended best practices for assessing true model generalizability (2) To Address Data Heterogeneity and Imbalance using advanced hybrid imputation and augmentation approaches (3) To Utilize Advanced Ensemble Methods, specifically stacking and hybrid classification architectures and (4) To Incorporate Explainable AI (XAI) for model versatility, interpretability, and enhanced clinical trust.

The expected outcomes include: (1) A generalizable noise-resilient diabetes prediction model validated using multiple external datasets. (2) Mitigation of imputation bias through the proposed hybrid imputation architecture. (3) Deployment-ready stacked ensemble classifier with enhanced robustness and stability across heterogeneous populations. (4) High-fidelity explainability insights support clinical decision-making and stakeholder trust.

2. Literature Review

Hameed et al. (2025) investigate the issue of *robustness and generalizability* in predictive models by conducting comprehensive external validation. They develop a diabetes-prediction model trained on one cohort and then systematically test it across multiple external datasets

drawn from different geographic or clinical populations. Their results show that, while within-sample performance is high, the model's predictive accuracy degrades significantly on external cohorts due to shifts in feature distributions, measurement protocols, and population characteristics. By quantifying this performance drop and analyzing which features contribute most to model decay, the authors argue for the necessity of routine external validation in clinical predictive modeling. They conclude that without such validation, models risk being overfit to their derivation cohorts and failing in real-world deployment.

Mittal et al. (2025) focus on building *universal frameworks* for integrating AI tools into clinical settings, emphasizing *real-world validation*. They propose a structured pipeline that goes beyond algorithm development to include deployment readiness, clinician-in-the-loop testing, and continuous feedback mechanisms. Their framework includes steps for harmonizing data sources, designing performance metrics aligned with clinical utility (e.g., decision thresholds, calibration), and conducting prospective pilot studies in real-world clinical environments. In their validation phase, they deploy a diabetes risk–prediction tool in multiple healthcare centers, measuring not only accuracy but also usability, clinician trust, and integration barriers. Their findings reveal that even well-performing models often face substantial friction in practice — due to differences in workflow, EHR integration, and data quality — underscoring that algorithmic development alone does not guarantee clinical value.

Kiran et al. (2025) address important *gaps in generalizability* from a psychosocial and demographic perspective in type 2 diabetes mellitus (T2DM) prediction. Their study argues that many predictive models neglect the psychosocial context — such as stress, social support, and behavioral health — which can materially influence diabetes risk. They build a model combining standard clinical/tabular predictors (age, BMI, glucose, lipids) with psychosocial variables, and they validate it on a demographically diverse population. While their enriched model offers modest gains in predictive power, they report challenges: psychosocial data are often noisy, prone to missingness, and collected inconsistently; moreover, integrating these variables across cohorts can hurt model portability. They conclude that to improve generalizability, future work must standardize psychosocial measurement and test models across diverse social and cultural contexts.

Khokhar et al. (2025) contribute by emphasizing both *explainable AI (XAI)* and *external validation* in their diabetes-prediction modeling. They propose an ensemble model (e.g., gradient boosting + random forest) and apply interpretability methods (such as SHAP) to understand feature importance and prediction behavior. Crucially, they don't stop at internal cross-validation: they externally validate their model on data from a completely independent healthcare system, finding that while accuracy remains high, the relative importance of features shifts — some predictors become less predictive, others more so. Their discussion highlights how explainability can reveal these shifts and guide model recalibration when transferring to new populations. They argue that combining XAI with external validation is a promising path toward robust, trustworthy clinical AI.

Sadeghi et al. (2025) examine *data heterogeneity and class imbalance* — two major obstacles to building generalizable predictive models. They propose a hybrid solution involving **synthetic data generation** (e.g., via GANs or SMOTE-like methods) combined with *advanced imputation*

strategies to fill missing values in tabular clinical datasets. They apply their methodology to a diabetes dataset with skewed class distribution and variable missingness, showing that synthetic augmentation plus robust imputation can improve model stability and reduce overfitting. However, they also note that synthetic data may introduce biases if not carefully crafted, and that imputation techniques may obscure real-world variance when deployed elsewhere. Their work underscores that dealing with data heterogeneity is not just a modeling issue, but a data-preparation challenge critical for generalizability.

Kumar et al. (2025) develop a *hybrid classification model* using an ensemble of classifiers combined via **stacked generalization** to predict diabetes. Their approach likely involves base learners (e.g., logistic regression, decision trees, SVM) whose outputs are fed into a meta-learner, improving robustness and reducing overfitting. They test this model on standard tabular datasets and report strong performance (e.g., higher AUC, precision) compared to individual models. Nonetheless, they also discuss that while stacking enhances predictive power, it can hamper interpretability and may not generalize well without external validation: performance gains seen in cross-validation may not carry over to unseen populations or datasets with different distributions. Their conclusion emphasizes balancing model complexity and generalizability, recommending that future work validate such stacked models on heterogeneous external cohorts.

3. Methodology

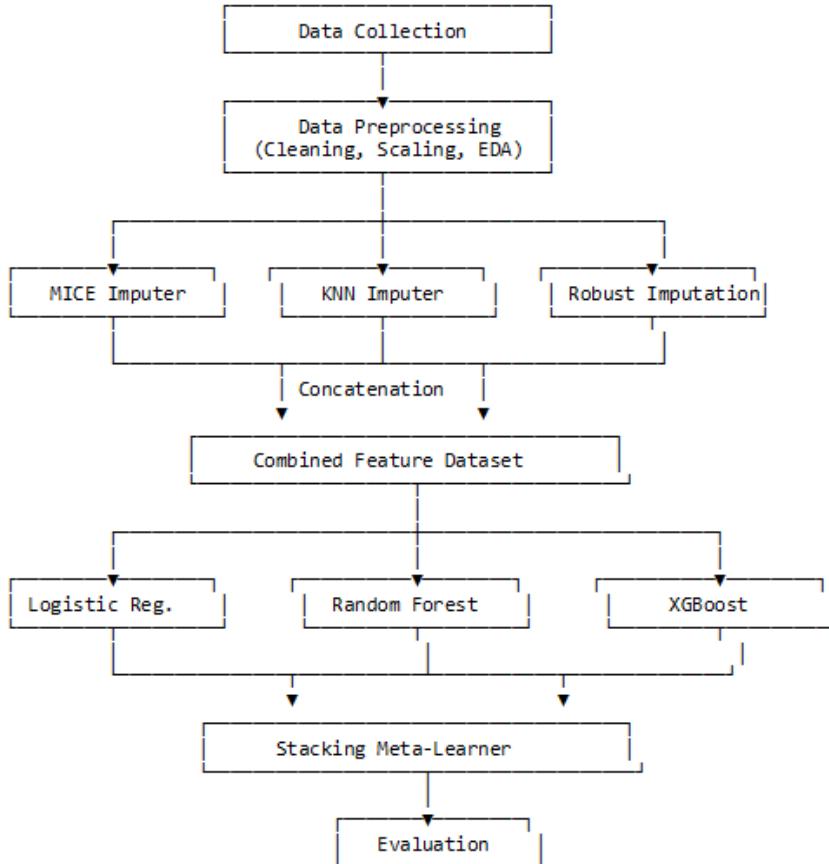


Figure 1: Methodology Flow Chart

Datasets include demographic, laboratory, and clinical features from multiple publicly available diabetes datasets (e.g., Pima Indians Diabetes Dataset, DaiBD dataset etc). External datasets ensure cross-cohort generalizability. **Data preprocessing steps:** (1) Exploratory Data Analysis (EDA), (2) Missingness mechanism analysis, (3) Outlier detection (IQR, z-score), (4) Encoding of categorical features and (5) Feature scaling (standardization). The proposed NR-HIE architecture comprises two core modules: (1) **Hybrid Imputation Streams** and (2) **Stacked Ensemble Classification**. Each method directly addresses the challenges and recommendations identified in the literature. (1) **Hybrid Imputation Module (Three Parallel Streams):** (a) **MICE (Multiple Imputations by Chained Equations):** It addresses multi-variable dependencies and complex missingness structures. This is recommended for heterogeneous clinical datasets. (b) **K-Nearest Neighbors Imputation:** It provides local, structure-aware filling of missing data suitable for varying patient clusters. It helps mitigate distributional imbalance by leveraging instance-level similarity. (c) **Robust/Constant Imputation (Median/Mode/Zero):** Stabilizes extreme noise regions and MNAR segments by injecting robust constant value. It ensures resilience under worst-case missingness patterns. The outputs of the three imputation pathways are fused using feature-space concatenation, mitigating the risk of **single-imputation-bias** and directly addressing the generalizability challenges highlighted. **Ensemble Learning Module:** The model incorporates three diverse base learners, each suited to a different clinical data property (1) **Logistic Regression (LR):** It provides high interpretability and linear signal decomposition. It serves as a baseline consistent with clinical modeling standards. (2) **Random Forest (RF):** It effectively handles nonlinearities, high variance, and mixed-type features. This contributes to the robustness. (3) **XGBoost:** It excels at modeling complex interactions and correcting residual error patterns. Enhances predictive dominance. **Stacked Generalization Meta-Learner:** For hybrid stacked architectures, predictions from the three base learners are fed into a **stacking meta-learner** (e.g., Logistic Regression or LightGBM) to: (a) Integrate diversified decision boundaries, (b) Reduce variance, (c) Improve cross-dataset consistency, and (d) Enhance generalizability across external validation cohorts.

4. Evaluation Metrics

The model will be evaluated using clinically preferred and statistically rigorous metrics.

Accuracy

Measures the proportion of correct predictions.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

Precision:

Indicates how many predicted positive cases are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

Measures how effectively the model identifies actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

Harmonic mean of precision and recall; suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Brier Score Loss:

$$BS = \frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2$$

The Wilcoxon Signed-Rank Test:

Used for paired statistical comparisons across imputation streams and ensemble variants.

$$z = \frac{T - \mu_T}{sd}$$

T = Smallest sum of rank (Positive or negative)

$$\mu_T = \frac{n(n+1)}{4}$$

SHAP Value Distribution (Qualitative)

The interpretability and feature contribution stability were assessed. It provides alignment with the XAI needs.

5. Tools and Technologies

Tools	Category	Task
Python	Programming Language	End-to-end implementation
VS code	IDE	To make notebook and working environment, documentation

Scikit-Learn	ML Framework	Imputation, preprocessing, base learners, stacking
XGBoost	Gradient Boosting Library	High-performance classifier
Matplotlib	Framework	Visualizations
Pandas	Data handling Framework	Data loading, wrangling, transformations, manipulations
NumPy	Framework	Numerical computations
SciPy	Statistical Computing	Wilcoxon tests, distributional analysis
Seaborn	Framework	Generate high quality statistical visualizations

6. 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,” *Int. Journal Med. Inform.*, 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.