Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 4(5) September 2026
DOI: https://dx.doi.org/10.4314/jobasr.v4i5.19

Application of Machine Learning Methods for Missing Data Imputation and Prediction of Diabetes Risk

Usman, Aliyu Goga[1]*, Yakubu Musa[1], Nasiru Sani Dauran[1] & Bello Arkilla Magaji[2]

[1]Department of Statistics, Faculty of Physical & Computing Sciences, Usmanu Danfodiyo University, Sokoto, Nigeria
[2]Department of Public Health, Usmanu Danfodiyo University, Sokoto, Nigeria
Corresponding author Email: usmanag@fudutsinma.edu.ng

ABSTRACT

Missing data is a near-universal feature of clinical datasets, yet its treatment is often relegated to a fixed preprocessing step that is not itself empirically validated. This study aimed to determine which machine learning regression engine most accurately imputes missing clinical values under different missing-data mechanisms, and whether the choice of imputer affects the accuracy of downstream diabetes risk prediction. Three versions of the same clinical dataset of 4,303 patients (Diabetes is the fourth leading cause of death in the world and one of the most common endocrine disorders. According to studies, Type 2 diabetes kills thousands of people around the world every year and imposes huge costs on societies in the form of surgeries and other treatment programs, as well as controlling complications and disability. Therefore, predicting and early diagnosis of this disease can greatly help governments and patients.

This dataset is the output of a Chinese research study conducted in 2018-24. It includes 1304 samples of patients who tested positive for diabetes, and the age of the participants ranges from 21 to 99 years old. The dataset was collected according to the indicators and standards of the World Health Organization, making it a reliable source for building diabetes diagnosis models. Researchers and healthcare professionals can use this dataset to train and test machine learning models to predict and diagnose diabetes in patients each containing 17 clinical and demographic predictors (age, gender, body mass index, blood pressure, glycaemic, lipid, liver and renal measures, and smoking, drinking and family history) and a binary diabetes outcome with a prevalence of approximately 30%, were each subjected to a different missing-data mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). A cross-file ground-truth reconstruction enabled direct, rather than proxy-based, evaluation of imputation accuracy on both synthetic and real missing cells. Decision Tree, Random Forest, and a linear-kernel Support Vector Regressor (SVR) were compared as the regression engine within an iterative (MICE-style) imputer using mean squared error (MSE), mean absolute error (MAE), and R^2, under 10–40% synthetic missingness and on real missing cells. Decision Tree, Random Forest, and Support Vector Machine (SVM) classifiers were then trained on the imputed datasets (stratified 80/20 train–test split, with 5-fold stratified cross-validation as a robustness check) and evaluated using accuracy, precision, recall (sensitivity), F1-score, and ROC-AUC. No formal statistical significance tests were performed, so differences between models were interpreted descriptively. Random Forest achieved the

best average imputation performance in the synthetic experiment (MSE = 76.48, MAE = 4.12, R^2 = 0.948), narrowly ahead of SVR (R^2 = 0.947) and clearly ahead of the Decision Tree (R^2 = 0.934), and its R^2 on real missing cells ranged from 0.954 to 0.959 across the three mechanisms (MAE 3.45–3.87). Random Forest and SVM classifiers showed very similar performance (ROC-AUC 0.911 and 0.912; accuracy 0.865 and 0.864; sensitivity 0.731 and 0.720; precision 0.806 and 0.809), both outperforming a single Decision Tree (ROC-AUC 0.868; accuracy 0.845; sensitivity 0.714), and performance was stable across the three missingness mechanisms. These results indicate that a well-chosen, model-based imputer can recover missing clinical values accurately and preserve downstream predictive performance, and that Random Forest is a suitable candidate for both imputation and classification in this dataset. However, the findings rest on a single clinical dataset with a reconstructed ground truth, and external validation in other populations is required before they can be generalised.

Keywords:
Missing data imputation; diabetes risk prediction; Random Forest; machine learning; MCAR; MAR; MNAR; clinical data

INTRODUCTION

Diabetes mellitus remains one of the most pressing chronic disease burdens worldwide. The International Diabetes Federation's 11th edition Diabetes Atlas estimates that approximately 589 million adults aged 20–79 were living with diabetes in 2024 (roughly one in nine of the global adult population), with more than four in ten cases remaining undiagnosed, and projects this figure will rise to 853 million by 2050 (International Diabetes Federation, 2025). The overwhelming majority of these cases are type 2 diabetes, a form strongly linked to modifiable risk factors including obesity, physical inactivity, and components of the metabolic syndrome (International Diabetes Federation, 2025). Given that a substantial proportion of affected individuals remain undiagnosed until complications emerge, there is a clear imperative to develop reliable, data-driven tools capable of identifying at-risk individuals from routinely collected clinical measurements (Jiang et al., 2023).

Machine learning algorithms such as Random Forest, Support Vector Machines, and Decision Trees have repeatedly demonstrated strong predictive capability for diabetes risk when trained on structured clinical variables (Ibitoye et al., 2024; Kumar et al., 2026). However, the practical utility of any such model is only as reliable as the completeness and integrity of the data on which it is trained, a condition rarely met in real-world clinical datasets. Missing data can arise from incomplete data collection, non-response, loss to follow-up, failure to attend clinical appointments, lack of measurements, and other deficiencies in the data collection process (Bell et al., 2019). And its consequences vary by mechanism: data may be Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR), each carrying different implications for the validity of downstream analysis (Little & Rubin, 2019). Naive handling through complete-case deletion or mean/median substitution risks introducing bias and degrading model performance, particularly when the missingness mechanism departs from MCAR (Aboelela et al., 2019).

Under Rubin's foundational taxonomy, the three mechanisms are distinguished as follows: MCAR, where missingness is unrelated to any observed or unobserved variable; MAR, where missingness depends on observed but not unobserved values; and MNAR, where missingness depends on the unobserved value itself (Little & Rubin, 2019). In clinical datasets, missingness rarely conforms to MCAR; it more often reflects systematic processes such as selective test ordering or differential access to diagnostic services, pointing toward MAR or MNAR mechanisms (Cardoso et al., 2024).

Traditional single-imputation methods (mean, median, or mode substitution, and hot-deck approaches) are computationally simple but are widely criticised for artificially reducing variance and distorting inter-variable relationships (Aubaidan et al., 2024). Multiple Imputation by Chained Equations (MICE) addresses some of these shortcomings by generating several plausible imputed datasets that reflect the uncertainty inherent in imputation (Azur et al., 2020). More recently, machine learning-based imputation methods (k-nearest neighbours, random forest-based approaches such as MissForest, and deep generative methods including denoising autoencoders and generative adversarial imputation networks) have gained prominence for their ability to model non-linear relationships among variables (Ishaq et al., 2024; Beaulieu-Jones & Moore, 2017).

Model-based imputation methods, which treat each missing variable as a regression problem conditioned on the remaining observed variables, have gained prominence as more robust alternatives. Random Forest-based imputers in particular have repeatedly outperformed classical statistical imputation methods on structured clinical data exhibiting non-linear inter-variable relationships (Hong & Lynn, 2020).

A wide range of machine learning algorithms has been applied to diabetes risk prediction, spanning logistic regression, decision trees, and support vector machines through ensemble and boosting methods such as Random Forest and XGBoost, as well as neural-network and deep-learning approaches capable of modelling complex patterns in patient data (Wu et al., 2021; El Massari et al., 2022).

Ensemble methods have generally shown strong performance compared with single classifiers in benchmark comparisons owing to their ability to combine multiple weak learners into a more robust composite model (Swaroop et al., 2024). Several studies have evaluated imputation strategies specifically within clinical or diabetes-related datasets (Chen & Chang, 2024; Damilola et al., 2024; Grigoroff et al., 2025).

Despite this progress, a clear research gap remains. Imputation accuracy, measured against a known ground truth, does not automatically guarantee strong downstream classification performance, so the two must be evaluated as related but distinct components of model performance (Shadbahr et al., 2023). Yet comparatively few studies have evaluated imputation accuracy and downstream classification performance jointly, across multiple missingness mechanisms, using a reconstructed ground truth for real (rather than only synthetically masked) missing cells.

This study addresses that gap by treating imputation and prediction as two linked stages of a single pipeline: the dataset completed by each candidate imputer becomes the direct input of the classifiers, so that the quality of the imputation can be traced through to the quality of the resulting diabetes predictions. Decision Tree, Random Forest, and Support Vector Regression (SVR) were selected as imputation engines because they represent three distinct and widely used learning paradigms: a single recursive-partitioning model that is simple but high-variance; a bagged ensemble of trees that reduces variance and captures non-linear interactions; and a linear-kernel, margin-based regressor with an ε-insensitive loss that provides a contrasting, near-linear benchmark. Each has a direct classification counterpart (Decision Tree, Random Forest, and Support Vector Machine, SVM), so the same three algorithm families can be compared at both the imputation and the prediction stages. The analysis uses three versions of the same clinical dataset of 4,303 patients (Diabetes is among the leading causes of death worldwide and is one of the most common endocrine disorders. Type 2 diabetes causes many deaths every year and places a heavy burden on societies through surgery, long-term treatment programmes, and the management of complications and disability. Early prediction and diagnosis can therefore benefit both patients and governments.

The data come from a Chinese research study conducted between 2018 and 2024. According to the source documentation, the study includes patients with a positive diabetes diagnosis, and the participants' ages range from 21 to 99 years. The data were collected according to the indicators and standards of the World Health Organization, which makes them a suitable basis for building diabetes diagnosis models that researchers and healthcare professionals can train and test each containing 17 clinical and demographic predictors and a binary diabetes outcome, and each subjected to a different missing-data mechanism (MCAR, MAR, or MNAR).

The study contributes to the literature in three ways. First, it evaluates imputation accuracy and downstream diabetes prediction simultaneously within a single design. Second, it introduces a cross-file ground-truth reconstruction that allows imputation accuracy to be measured directly on real, rather than only synthetic, missing cells across all three missingness mechanisms. Third, it identifies and corrects a hidden sentinel-value artefact that would otherwise have biased both imputation and classification.

The specific objectives of this study are to: (i) characterise the nature and extent of missing data in a diabetes-related clinical dataset; (ii) compare Decision Tree, Random Forest, and Support Vector Regression (SVR)-based imputation methods; (iii) develop classification models for diabetes risk prediction on the resulting imputed datasets; (iv) evaluate predictive performance across imputation strategies using standard classification metrics; and (v) assess the stability of the best-performing imputation–prediction combination across missingness mechanisms and missingness levels.

MATERIALS AND METHOD

Description of the Dataset

Source of the Data

Diabetes is among the leading causes of death worldwide and is one of the most common endocrine disorders. Type 2 diabetes causes many deaths every year and places a heavy burden on societies through surgery,

long-term treatment programmes, and the management of complications and disability. Early prediction and diagnosis can therefore benefit both patients and governments.

The data come from a Chinese research study conducted between 2018 and 2024. According to the source documentation, the study includes patients with a positive diabetes diagnosis, and the participants' ages range from 21 to 99 years. The data were collected according to the indicators and standards of the World Health Organization, which makes them a suitable basis for building diabetes diagnosis models that researchers and healthcare professionals can train and test.

Structure of the Dataset

The study uses three versions of the same clinical dataset, named Diabetes_MCAR, Diabetes_MAR and Diabetes MNAR. Each version contains 4,303 patient records and 18 variables, giving 77,454 cells per file. The versions hold identical underlying records and differ only in the cells that have been set to missing under three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). No duplicate records were found in any version.

The 18 variables comprise 17 predictors and one binary outcome, Diabetes. The predictors are demographic (Age, Gender), anthropometric (BMI), haemodynamic (SBP, DBP), biochemical (FPG, Chol, Tri, HDL, LDL, ALT, BUN, CCR, FFPG) and lifestyle or history indicators (smoking, drinking, family_histroy). The outcome is complete in all three files.

Description of Variables

Table 1: Variables in the dataset

Variable	Description	Type (unit)	Observed range
Age	Age of the participant	Continuous (years)	22–93
Gender	Sex of the participant (coded 1 and 2)	Binary	1–2
BMI	Body mass index	Continuous (kg/m²)	15.6–45.8
SBP	Systolic blood pressure	Continuous (mmHg)	72–199
DBP	Diastolic blood pressure	Continuous (mmHg)	47–131
FPG	Baseline fasting plasma glucose	Continuous (mmol/L)	2.84–6.99
Chol	Total cholesterol	Continuous (mmol/L)	1.65–11.65
Tri	Triglycerides	Continuous (mmol/L)	0–32.64
HDL	High-density lipoprotein cholesterol	Continuous (mmol/L)	0–4.86
LDL	Low-density lipoprotein cholesterol	Continuous (mmol/L)	0.54–6.27
ALT	Alanine aminotransferase	Continuous (U/L)	4.5–436.2
BUN	Blood urea nitrogen	Continuous (mmol/L)	1.38–17.73
CCR	Creatinine clearance	Continuous	4.86–307
FFPG	Final fasting plasma glucose	Continuous (mmol/L)	3.2–29.7
smoking	Smoking status (1 = current, 2 = ever, 3 = never smoker)	Categorical (coded)	1–3
drinking	Drinking status (1 = current, 2 = ever, 3 = never drinker)	Categorical (coded)	1–3
family_histroy	Family history of diabetes (spelling of the column name as in the file)	Binary (0 = no, 1 = yes)	0–1
Diabetes	Diabetes outcome (target variable)	Binary (0 = no, 1 = yes)	0–1

Note: ranges are taken from the observed values in the MAR file. Smoking, drinking and family history codes

follow the source documentation. The coding of Gender (1 and 2) and the measurement units are not stated in the source description and should be confirmed.

Outcome Variable

The target variable Diabetes is binary. Of the 4,303 patients, 1,303 (30.3%) are coded 1 (diabetes) and 3,000 (69.7%) are coded 0 (no diabetes), so the classes are moderately imbalanced. The source documentation reports 1,304 positive cases, one more than the version analysed here. Patients with diabetes are older on average (59.7 versus 43.0 years), have a higher mean BMI (25.3 versus 23.6 kg/m²) and have a higher baseline FPG (5.83 versus 4.96 mmol/L) than those without. All baseline FPG values are below 7.0 mmol/L, which suggests that the cohort was free of diabetes at baseline and that the outcome reflects incident diabetes.

Missing Data Patterns

Table 2 summarizes the extent of missingness in each version.

Table 2: Summary of missing data by mechanism

Characteristic	MCAR	MAR	MNAR
Variables with missing values	17 predictors	10 (SBP, DBP, FPG, Chol, Tri, HDL, LDL, ALT, BUN, FFPG)	10 (Age, BMI, SBP, DBP, FPG, Chol, Tri, LDL, ALT, FFPG)
Missing per affected variable	19–21% (808–913 cells)	8.3–16.6% (357–713 cells)	1.5–20.8% (64–895 cells)
Overall missing cells	18.8%	5.8%	5.1%
Records with at least one missing value	4,218 (98.0%)	2,782 (64.7%)	2,625 (61.0%)
Completely observed variables	Diabetes	Age, Gender, BMI, CCR, smoking, drinking, family_histroy, Diabetes	Gender, HDL, BUN, CCR, smoking, drinking, family_histroy, Diabetes

MCAR. Missingness is spread evenly over all 17 predictors, at roughly 19–21% per variable. The missing rate does not depend on the outcome (19.7% for patients without diabetes and 20.4% for patients with diabetes).

MAR. Missingness depends on variables that remain observed. It is higher among patients with diabetes (for example ALT 25.2% versus 12.8%, and FPG 15.0% versus 9.1%) and increases with age (correlation of about 0.2 between age and the missingness indicators of SBP, DBP, FPG, BUN and FFPG).

MNAR. Missingness is designed to depend on the unobserved values themselves. The observed means in this version are lower than those in the MCAR version (for example FPG 5.17 versus 5.23 mmol/L, Age 47.0 versus 48.0 years and LDL 2.93 versus 2.98 mmol/L), which is consistent with larger values being more likely to be removed. The mechanism cannot be verified from the observed data alone.

Data Cleaning

Exploratory inspection revealed a recurring sentinel value (4.860753) contaminating multiple, clinically unrelated numeric columns (smoking, drinking, HDL, LDL, Chol, Tri, ALT, BUN, CCR), appearing up to eight times within a single patient row. Because the value was clinically implausible (e.g., as the minimum recorded Creatinine Clearance Rate) and recurred identically across unrelated variables, it was treated as a data-generation artefact and recoded as missing across all three files prior to further processing. This step increased average per-column missingness (e.g., from 18.81% to 21.64% in the MCAR file) but removed a confound that would otherwise have biased both imputation and downstream classification.

Ground-Truth Reconstruction

Because the three files share an identical patient cohort, a reference ("ground-truth") table was constructed by taking, for each cell, the first non-null value encountered across the MCAR, MAR, and MNAR files post-cleanup. This recovered all but a small residual (<1% of cells, concentrated in HDL, LDL, BUN, smoking, and drinking) where all three mechanisms coincided in deleting the same patient's value; these residual gaps were filled with the column median. This reconstructed table served as the evaluation reference for all imputation accuracy metrics reported in the Results section.

Feature Engineering

Features of Dataset: Age Gender BMI SBP (Systolic Blood Pressure) DBP (Diastolic Blood Pressure) FPG (Fasting Plasma Glucose) FFPG (Final Fasting Plasma Glucose) Cholesterol Triglyceride HDL (High-Density

Lipoprotein) LDL (Low-Density Lipoprotein) ALT (Alanine Aminotransferase) BUN (Blood urea nitrogen) CCR (Creatinine Clearance) Smoking Status: (1: Current Smoker, 2: Ever Smoker, 3: Never Smoker) Drinking Status: (1: Current Drinker, 2: Ever Drinker, 3: Never Drinker) Family History of Diabetes: (1: Yes, 0: No) Diabetes

Imputation Model Formulation

Three estimator families were evaluated as the regression engine within scikit-learn's Iterative Imputer, which imputes each column in turn as a supervised regression problem conditioned on all other columns, cycling until convergence (a MICE-style multivariate imputation procedure): a Decision Tree Regressor, which recursively partitions the predictor space by minimising intra-node variance; a Random Forest Regressor, an ensemble of bootstrap-trained trees with random feature subsampling whose predictions are averaged to reduce variance; and a linear-kernel Support Vector Regressor (SVR), which fits a hyperplane minimising an ε-insensitive loss. Model selection was performed on the reconstructed ground-truth table rather than on the real missing cells, to allow controlled comparison: synthetic MCAR-style missingness was injected independently into each of 13 continuous clinical columns at four levels (10%, 20%, 30%, 40%), and each candidate imputer's predictions at the masked cells were compared to the known ground truth using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Throughout this paper, SVR denotes the Support Vector Regressor used for continuous-value imputation, whereas SVM denotes the Support Vector Machine classifier used for diabetes prediction.

Model Selection

Model selection is a critical step in the development of machine learning (ML) solutions for data imputation, particularly in the context of health data integrity. The choice of model significantly impacts the accuracy, robustness, and generalizability of the imputation Random Forest Imputation

Random Forest Imputation is a powerful method for handling missing data using Random Forest algorithms. A Random Forest is an ensemble learning method that combines multiple decision trees to improve the predictive accuracy and control over-fitting. In the context of imputation, a Random Forest model learns the relationship between the features and uses this to predict and fill in missing values.

Model Representation

Let $T_b(X_{obs})$ be the b-th decision tree in the forest, where b=1, 2….,B.

The Random Forest prediction is:

$$\hat{Y}_{Miss} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(X_{obs}) \qquad (3)$$

process. This section explores the criteria and methods used to select appropriate ML models for imputing missing data in healthcare.

Decision Tree Imputation

Decision Tree Imputation is a technique for filling in missing data using decision trees. A decision tree is a type of model that splits the data into branches to make predictions based on the values of input features. When used for imputation, the decision tree learns how each feature relates to the others and predicts the missing values based on this learned structure.

Model Representation

let $X = \{X_1, X_2, \ldots, X_n\}$ represent the set of features in the datasets

1. Y_{miss}: the target variable with missing values.
2. X_{obs}: the observed features (variable without missing values).

A decision tree model $T(X_{obs})$ is constructed using the observed data:

$$\hat{Y}_{miss}=T(X_{obs}) \qquad (1)$$

Algorithm Steps

1. Partition the dataset based on the observed features X_{obs} .
2. At each node, select the feature X_j that maximizes the information gain or minimizes the impurity (e.g., Gini index or entropy).
3. Repeat the partitioning until each terminal node (leaf) contains a small subset of data.
4. Predict the missing value Y_{miss} at each leaf node using the average or majority class of the observed values.

Imputation

Given a decision tree (X_{obs}) , the predicted missing value is:

$$\hat{Y}_{miss}= \frac{1}{N_l}\sum_{i\in l} Y_i \quad (2)$$

Where N_l is the number of samples in the leaf node *L* and Y_i are the observed values in that node.

Algorithm Steps

1. Generate *B* bootstrap samples from the original dataset.
2. Construct a decision tree $T_b(X_{obs})$ for each bootstrap sample, using a random subset of features at each split.
3. Aggregate the predictions from all *B* trees to obtain the final imputation.

The imputed value is the average of the predictions from all trees:

$$\hat{Y}_{Miss} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_{lb}}\sum_{i\in l_b} Y_i \qquad (4)$$

Where L_b is the leaf node in the b-th tree containing the sample, and N_{lb} is the number of samples in that node.

Support Vector Regression (SVR) Imputation is a technique for handling missing data using Support Vector Regression, the regression form of the Support Vector Machine (SVM). SVR is a supervised machine learning algorithm that fits a function to a continuous target while tolerating deviations within an ε-insensitive margin. When used for imputation, an SVR model learns the relationships between features and predicts missing values based on this learned structure

Model Representation:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i max\left(0, \left|Y_i - w^T\emptyset\left(X_{obs,i}\right) - b - \in\right)\right. \quad (6)$$

where C is the regularization parameter and ϵ controls the width of the ϵ-insensitive zone.

The missing value is imputed as:

$$\hat{Y}_{Miss} = \sum \alpha_i K\left(X_{Obs,i}, X_{Obs}\right) + b \quad (7)$$

Where α_i are the dual coefficients and $K(\cdot,\cdot)$ is the kernel function.

Classification Model Formulation and Evaluation Protocol

Using the fully imputed, feature-engineered datasets, three classifier families (Decision Tree, Random Forest, and a Support Vector Machine (SVM) classifier with a maximum-margin separating hyperplane) were trained to predict the binary Diabetes outcome, independently for each missingness mechanism. For each mechanism/model combination, the dataset was split into 80% training and 20% testing subsets using stratified sampling to preserve class balance, and performance was assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC on the held-out test set. To verify that results were not an artefact of a single split, 5-fold stratified cross-validation was additionally run on the full dataset for each combination. Feature importance for the Random Forest classifier was extracted using the model's built-in impurity-based importance measure.

In classification tasks, the goal is to assign input data to one of several predefined categories. Common performance metrics for classification include:

Accuracy:

Accuracy is the proportion of correct predictions out of the total number of predictions.

Let X_{Obs} be the input features and Y_{Miss} the target missing value.

The SVR model finds a function $f(X_{obs})$ that approximates the missing values:

$$\hat{Y}_{Miss} = f(X_{obs}) = W^T\emptyset(X_{obs}) + b \quad (5)$$

Where $\emptyset(\cdot)$ is a kernel function mapping the input features to a higher-dimensional space, w are the model weights, and b is the bias term.

The SVR model solves the following optimization problem:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

where:

i. TP (True Positives): Correctly predicted positive cases.
ii. TN (True Negatives): Correctly predicted negative cases.
iii. FP (False Positives): Incorrectly predicted as positive.
iv. FN (False Negatives): Incorrectly predicted as negative.

Precision:

Precision measures the accuracy of positive predictions, i.e., the proportion of true positive predictions out of all positive predictions.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall (Sensitivity or True Positive Rate):

Recall is the proportion of actual positive cases that were correctly identified by the model.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1 Score:

The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

Use Case: The F1 Score is particularly useful when the dataset is imbalanced, as it balances the trade-off between precision and recall.

Confusion Matrix:

The confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positives, true negatives, false positives, and false negatives.

Structure:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

RESULTS

Data Cleaning and Missingness Profile

Table 1 summarises the scale of the sentinel-value cleanup. The native missingness rates differed substantially by mechanism (roughly 19–22% per column for MCAR versus 5–9% for MAR and MNAR), consistent with MCAR being generated as uniform random deletion across all columns, while MAR and MNAR concentrate missingness within specific clinical variables such as the lipid panel and renal-function columns.

Table 3. Data cleaning summary across the three missingness mechanisms.

Dataset	Rows × Cols	Sentinel Values Removed	Avg. Missing % (raw)	Avg. Missing % (after cleanup)
MCAR	4,303 × 18	2,190	18.81	21.64
MAR	4,303 × 18	2,670	5.78	9.22
MNAR	4,303 × 18	2,656	5.05	8.48

Imputation Model Selection (10–40% Synthetic Missingness)

Table 3 reports average imputation performance across the four synthetic missingness levels. Random Forest achieved the highest average R^2 ($R^2 = 0.948$), narrowly ahead of SVR ($R^2 = 0.947$) and clearly ahead of the single Decision Tree ($R^2 = 0.934$). The margin between Random Forest and SVR is very small ($\Delta R^2 = 0.001$), and because no formal statistical comparison was performed, this ranking is descriptive and should not be taken as evidence of statistically meaningful superiority. All three models degraded gracefully as missingness increased from 10% to 40%, with R^2 declining from approximately 0.95–0.96 to 0.92–0.94. These results indicate that substantial predictive information remained in the observed variables, allowing relatively accurate regression-based imputation even at higher levels of synthetic missingness.

Table 4. Average imputation performance across the 10–40% synthetic missingness range, ranked by R^2.

Model	MSE (avg)	MAE (avg)	R^2 (avg)
Random Forest	76.482	4.121	0.948
SVR	78.328	4.147	0.947
Decision Tree	97.559	4.505	0.934

Imputation on Real Missing Data

Table 4 reports imputation accuracy on the actual missing cells of each file (native missingness of 21.3% for MCAR, 9.6% for MAR, and 8.6% for MNAR after cleanup), validated against the reconstructed ground truth. Random Forest achieved the highest R^2 for MCAR and MNAR and performed comparably with SVR for MAR, confirming its consistently strong performance across the three mechanisms. Its performance was marginally higher on the MAR and MNAR datasets than on the MCAR dataset, which may partly reflect their lower overall missingness and the concentration of missing values within correlated clinical variables.

Table 5. Imputation accuracy on real missing cells, validated against reconstructed ground truth.

Mechanism	Model	Missing %	MSE	MAE	R^2
MCAR	Decision Tree	21.29	90.771	4.217	0.937
MCAR	Random Forest	21.29	66.071	3.869	0.954
MCAR	SVR	21.29	70.832	3.880	0.951
MAR	Decision Tree	9.59	94.261	3.764	0.938
MAR	Random Forest	9.59	63.075	3.445	0.959
MAR	SVR	9.59	62.386	3.268	0.959
MNAR	Decision Tree	8.56	95.316	3.905	0.946
MNAR	Random Forest	8.56	80.471	3.587	0.955
MNAR	SVR	8.56	88.192	3.715	0.950

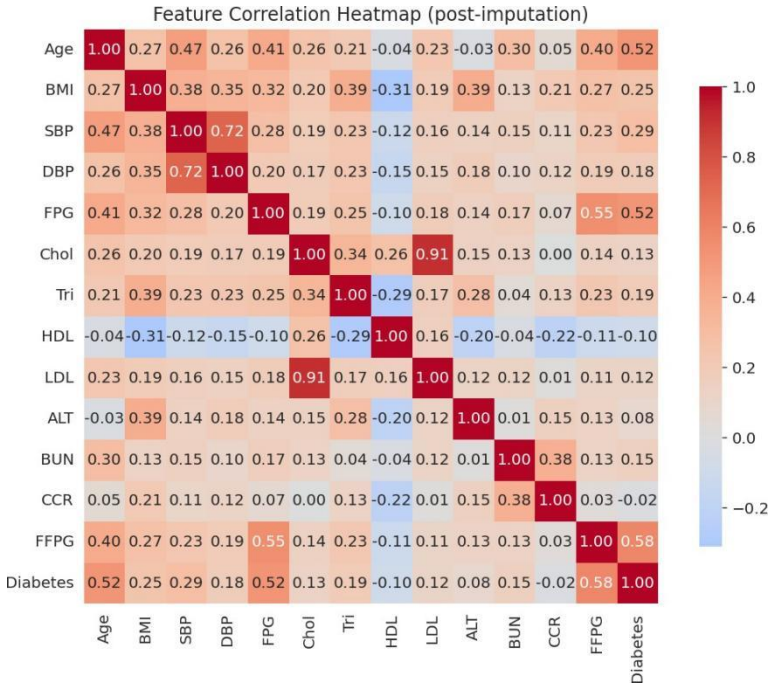


Figure 1. Feature correlation heatmap computed on the fully imputed MCAR dataset.

Diabetes Classification Performance

Table 5 reports held-out test-set classification performance, averaged across all three missingness mechanisms. Random Forest and SVM showed very similar overall performance, with ROC-AUC values of 0.911 and 0.912, respectively, with Random Forest holding a small edge in Recall and F1, and SVM a small edge in Precision. The Decision Tree was clearly the weakest model on every metric, roughly 4–5 ROC-AUC points below the other two. Performance was remarkably stable across the three missingness mechanisms, with no model's ROC-AUC varying by more than about one point between MCAR, MAR, and MNAR evidence that the Random Forest imputation step preserved the predictive signal needed for diabetes classification regardless of the original missingness mechanism. Five-fold cross-validation results closely tracked the held-out test metrics for all nine mechanism/model combinations, with standard deviations no larger than 0.014.

Table 6. Classification performance averaged across all three missingness mechanisms (overall ranking).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
SVM	0.864	0.809	0.720	0.762	0.912
Random Forest	0.865	0.806	0.731	0.766	0.911
Decision Tree	0.845	0.763	0.714	0.736	0.868

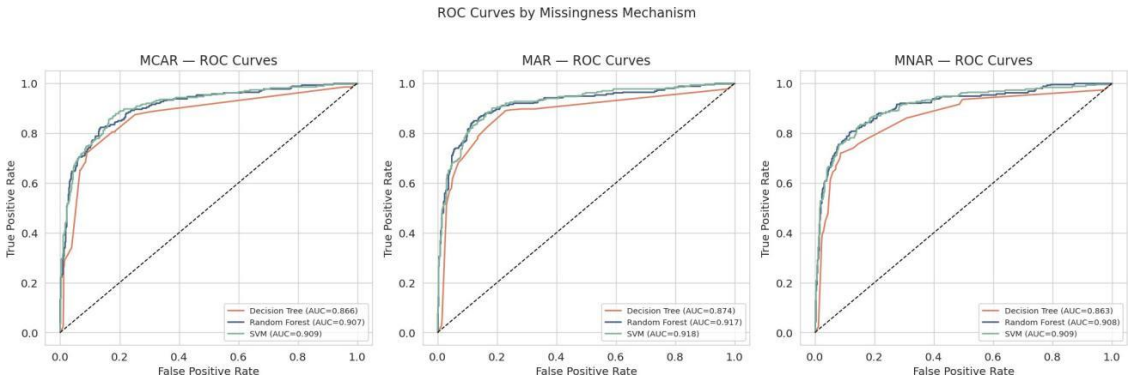


Figure 2. ROC curves for each classifier, plotted separately for the MCAR, MAR, and MNAR imputed test sets.

A consistent asymmetry was observed across all nine mechanism×model combinations: models generated more false negatives (missed diabetic cases) than false positives, reflecting the approximately 30% class prevalence of the outcome rather than a flaw specific to any one algorithm. If the clinical priority is to avoid missing true diabetic cases, Random Forest's slightly higher recall makes it the marginally preferable choice; if minimising false alarms is prioritised, SVM's slightly higher precision would be preferred.

Feature Importance

Table 5 reports the top predictors from the Random Forest classifier's impurity-based feature importance. The two fasting-glucose measurements (FFPG and FPG) had the highest importance values, together accounting for nearly half of the total importance, followed by Age and the engineered Age Group. These findings indicate that glucose-related variables and age contributed substantially to the Random Forest model's predictions. Engineered features (Pulse Pressure and Age Group) also appeared among the top-ranked predictors, suggesting that these derived variables contributed to model prediction, whereas BMI Category and Gender ranked low (Figure 3).

Table 5. Top ten feature importances from the Random Forest classifier.

Feature	Importance
FFPG	0.2776
FPG	0.1891
Age	0.1118
Age_Group	0.0763
SBP	0.0366
Pulse_Pressure	0.0327
BMI	0.0300
Tri	0.0280
HDL	0.0263
Chol	0.0260

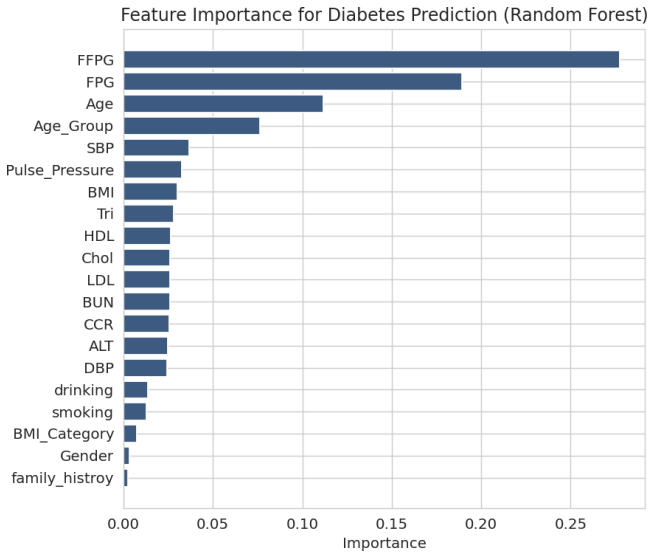


Figure 3. Full feature importance ranking (Random Forest classifier).

DISCUSSION

The consistency with which Random Forest showed the strongest overall imputation performance relative to both the Decision Tree and the linear-kernel SVR across the synthetic missingness gradient and across the three real missingness mechanisms can be interpreted through a bias–variance lens. A single Decision Tree used inside an iterative imputer is a relatively high-variance learner: small perturbations in which rows are complete at a given imputation round produce materially different partitions, and these errors compound across iterations. Random Forest's bagging-plus-feature-subsampling mechanism can reduce this variance, consistent with the observed ranking. The near-tie between Random Forest and the linear SVR is nonetheless informative: because a linear kernel can only recover linear or near-linear relationships, its competitiveness implies that a substantial share of the exploitable inter-variable structure in this dataset (physiologically paired variables such as SBP–DBP and FPG–FFPG) is approximately linear, with Random Forest's advantage concentrated in the residual non-linear component.

A theoretically notable result is that Random Forest performed marginally better on the real MAR and MNAR files than on the real MCAR file, despite MCAR being the mechanism classically regarded as least problematic under Rubin's taxonomy. This inversion is plausibly explained by the fact that MAR and MNAR missingness in this dataset is concentrated within a correlated sub-block of variables (the lipid panel and renal markers), leaving more columns simultaneously observed within any given row (precisely the conditioning information a multivariate imputer depends on), whereas MCAR spreads missingness diffusely and can remove several mutually informative columns from the same row simultaneously. This qualifies Rubin's framework as applied in a machine-learning context: under a model-based, multivariate imputer, the informativeness of the missingness pattern can matter more than its formal mechanism classification in determining ease of reconstruction, even though the classification remains essential for reasoning about the statistical validity of the resulting estimates.

The dissociation between imputation accuracy and downstream classification performance is a further substantive finding. Random Forest and SVM classifiers converged to very similar ROC-AUC regardless of which mechanism's imputed dataset they were trained on, even though imputation R^2 varied measurably by mechanism. This supports the view that a sufficiently accurate imputer can neutralise much of the theoretical distinction between missingness mechanisms at the point of downstream inference, and suggests a threshold effect in which further gains in reconstruction accuracy beyond a certain fidelity yield diminishing returns for the classifier sitting downstream of it.

Two methodological features of the design merit emphasis. First, the cross-file ground-truth reconstruction provides a genuine reference against which imputation on real, rather than only synthetically masked, missing cells can be scored, a comparison many studies in the literature are unable to make. Second, the sentinel-value discovery demonstrates that data cleaning functioned as an active analytical contribution rather than a perfunctory step, since an undetected artefact of that scale would have silently biased both imputation and downstream classification. A limitation that should be stated candidly is that the ground-truth table itself is a

reconstruction: accurate for the great majority of cells recoverable from cross-file overlap, but filled with column medians for a small residual where all three mechanisms coincided in deleting the same value. This does not undermine the study's comparative conclusions but introduces a mild source of optimism into the absolute accuracy reported for the affected biomarkers. A further limitation is that the iterative imputation and the mode imputation of categorical variables were applied to each complete dataset before the train–test split, so that test-set records contributed to the fitting of the imputation models and the reported classification performance may be mildly optimistic; future work should fit the imputer within the training folds only. In addition, all findings derive from a single clinical dataset, and comparisons among models were descriptive because no formal statistical tests were performed.

CONCLUSION AND RECOMMENDATIONS

This study evaluated a complete machine-learning-based pipeline for handling real-world clinical missing data: detecting a hidden data-quality issue, reconstructing a reference dataset from aligned files, selecting an imputation model through controlled evaluation across a 10–40% synthetic missingness range, and confirming that the resulting imputed data supports accurate and stable downstream diabetes prediction. Random Forest demonstrated strong and consistent performance as an imputation method and achieved classification performance comparable to SVM for this dataset and task. The findings therefore support Random Forest as a suitable candidate for both imputation and classification in the present dataset, although external validation is required before generalising this conclusion to other clinical populations or settings. The choice between Random Forest and SVM as a classifier may depend on the relative importance assigned to Recall and Precision in the intended application.

More broadly, the study supports treating missing-data handling in health research as a data-reliability problem rather than merely a data-completion problem: an appropriate imputation procedure should demonstrate acceptable numerical accuracy, robustness to different levels and mechanisms of missingness, preservation of clinically meaningful statistical relationships, and continued usefulness for downstream analysis. Future work should extend this framework to non-linear kernel variants of the SVR, larger and more heterogeneous clinical cohorts, and deep generative imputation methods, to test whether the threshold effect identified here between imputation accuracy and classification performance generalises beyond the present setting.

REFERENCES

Aboelela, A., Feric, Z., & Kaeli, D. (2019). *Using machine learning for data imputation to support environment health studies in Northern Puerto Rico*. 2019 IEEE MIT Undergraduate Research Technology Conference (URTC). https://doi.org/10.1109/URTC49097.2019.9660568

Aubaidan, B. H., Kadir, R. A., & Ijab, M. T. (2024). A comparative analysis of SMOTE and CSSF techniques for diabetes classification using imbalanced data. *Journal of Computer Science, 20*(9), 1146–1165. https://doi.org/10.3844/jcssp.2024.1146.1165

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329

Beaulieu-Jones, B. K., & Moore, J. H. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Biocomputing 2017: Proceedings of the Pacific Symposium* (pp. 207–218). World Scientific. https://doi.org/10.1142/9789813207813_0021

Bell, M. L., Fairclough, D. L., Fiero, M. H., & Butow, P. N. (2019). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology, 11*, 857–867. https://doi.org/10.2147/CLEP.S173884

Cardoso, P., Dennis, J. M., Bowden, J., Shields, B. M., & Trevelyan, J. (2024). Dirichlet process mixture models to impute missing predictor data in counterfactual prediction models: An application to predict optimal type 2 diabetes therapy. *BMC Medical Informatics and Decision Making, 24*, Article 12. https://doi.org/10.1186/s12911-023-02400-3

Chen, C.-Y., & Chang, Y.-W. (2024). Missing data imputation using classification and regression trees. *PeerJ Computer Science, 10*, e2119. https://doi.org/10.7717/peerj-cs.2119

El Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization, 10*(2), 319–338. https://doi.org/10.13052/jicts2245-800X.10212

El Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT*

Standardization, *10*(2), 319–337. https://doi.org/10.13052/jicts2245-800X.10212

Grigoroff, L., Masuda, R., Lindon, J., Kadyrov, J., Nicholson, J. K., Holmes, E., & Wist, J. (2025). Evaluation of imputation strategies for multi-centre studies: Application to a large clinical pathology dataset. *PLOS ONE*, *20*(11), e0335852. https://doi.org/10.1371/journal.pone.0335852

hadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Viñas Torné, R., Sala, E., & Schönlieb, C.-B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, *3*, 139. https://doi.org/10.1038/s43856-023-00356-z

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*, 199. https://doi.org/10.1186/s12874-020-01080-1.

Ibitoye, A. O. J., Akinyemi, J. D., & Onifade, O. F. W. (2024). Machine learning-based diabetes risk prediction using associated behavioral features. *Computing Open*. https://doi.org/10.1142/S2972370124500065.

Ibitoye, A. O. J., Akinyemi, J. D., & Onifade, O. F. W. (2024). Machine learning-based diabetes risk prediction using associated behavioral features. *Computing Open*, *2*. https://doi.org/10.1142/S2972370124500065

International Diabetes Federation. (2025). *IDF diabetes atlas* (11th ed.). International Diabetes Federation. https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

Ishaq, M., Zahir, S., Iftikhar, L., Bulbul, M. F., Rho, S., & Lee, M. Y. (2024). Machine learning based missing data imputation in categorical datasets. *IEEE Access, 12*, 88332–88344. https://doi.org/10.1109/ACCESS.2024.3411817

Jiang, L., Xia, Z., Zhu, R., Gong, H., Wang, J., Li, J., & Wang, L. (2023). Diabetes risk prediction model based on community follow-up data using machine learning. *Preventive Medicine Reports*, *35*, 102358. https://doi.org/10.1016/j.pmedr.2023.102358

Kumar, N., Bhattacharjee, A., Vishwakarma, G. K., & Majumdar, T. (2026). MIGEE: A comprehensive package for missing data imputation and longitudinal analysis in large-scale datasets. *Journal of Computational Science*, *99*, 102906. https://doi.org/10.1016/j.jocs.2026.102906.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Swaroop, C. R., Jayamanasa, V., Shankar, R. S., Babu, M. G., Shariff, V., & Kumar, N. S. K. M. (2024). Optimizing diabetes prediction through intelligent feature selection: A comparative analysis of Grey Wolf Optimization with AdaBoost and Ant Colony Optimization with XGBoost. In *Algorithms in Advanced Artificial Intelligence: ICAAAI-2023* (pp. 311–318). CRC Press. https://doi.org/10.1201/9781003529231-47

Wu, Y., Hu, H., Cai, J., Chen, R., Zuo, X., Cheng, H., & Yan, D. (2021). Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Frontiers in Public Health*, *9*, 626331. https://doi.org/10.3389/fpubh.2021.626331