Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19103-0

Impact of machine learning-based imputation techniques on medical datasets- a comparative analysis

Shweta Tiwaskar[1] · **Mamoon Rashid**[1] · **Prasad Gokhale**[1]

Received: 1 June 2023 / Revised: 1 February 2024 / Accepted: 27 March 2024

Abstract
In the realm of medical datasets, particularly when considering diabetes, the occurrence of data incompleteness is a prevalent issue. Unveiling valuable patterns through medical data analysis is crucial for early and precise medical predictions. However, the quality of data and the proper handling of missing data hold significant significance. To address this challenge, imputation stands as a robust approach. The main goal of this paper aims to provide a comprehensive investigation into the effects brought about by Machine Learning (ML) based imputation techniques, specifically K Nearest Neighbor Imputation (KNNI), Multiple Imputation by Chained Equations (MICE), and MissForest. Results of all three techniques are compared with the complete dataset for five missing rates (10% to 50%), and evaluated using four categories of evaluation criteria i.e. (1) model performance, (2) imputation error rate (Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2) values), (3) Pearson correlation analysis and, (4) model selection basis (Bayesian information criterion (BIC), Akaike information criterion (AIC), values). Model performance includes accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (Mcoff) score of four ML classifiers viz. (a) Random Forest (RF), (b) Support vector machine (SVM), (c) AdaBoost, (d) XGBoost (XGB). For all missing rate cases, the MissForest technique is better than the KNNI and MICE in accuracy and Mcoff in 80% of cases, precision in 40% of cases, recall in 60% of cases, F1 score, MAE, RMSE, R^2 in 100% of cases, AIC in 80% of cases, and BIC values in 100% of cases. Also, the correlation analysis confirms that the MissForest imputation preserves association between the variables, like the complete dataset. Overall, our findings suggest that MissForest is a better machine learning-based imputation technique for handling missing data in diabetes research.

Keywords Missing data · Machine learning-based imputation techniques · Diabetes prediction

✉ Mamoon Rashid
mamoon873@gmail.com

[1] Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune, India

1 Introduction

According to IDF (International Diabetes Federation) Diabetes Atlas, the worldwide occurrence rate of diabetes in the 20 to 79 age group was 10.5% in 2021, equivalent to approximately 536.6 million people. This figure is projected to increase to 12.2% by 2045, reaching around 783.2 million individuals [1]. Diabetes is known to be associated with severe complications such as retinopathy, neuropathy, cancer, heart attacks, and potential fatality [2, 3]. The high prevalence of diabetes, and the absence of intelligent techniques, cause delays and inaccuracies in the process of diagnosis. Medical data mining has the capability to uncover concealed patterns from vast amounts of data, leading to timely and precise medical decisions [4, 5]. This can be applied to accurate diabetes prediction as well if sufficient and quality data is available. Data quality and missing data are common problems, with real-world diabetes datasets, which affect the performance of intelligent techniques [6, 7]. In the healthcare domain, patient records are frequently produced as a result of patient care activities, rather than being explicitly collected as part of a structured research protocol, resulting in the potential loss of valuable information [8]. Hence, a significant portion of patient records exhibit missing values, as evidenced by the presence of datasets in the UCI (University of California Irvine) Machine Learning Repository that contain more than 40% missing values [9]. There are many reasons for missingness in medical data such as unrecorded values, incorrect measurements, equipment errors, human errors, outliers, or wrong data. In handling missingness, it's important to know the missingness mechanism, or cause of missingness, and the missingness pattern, which can impact the choice of imputation techniques [10]. The existing literature categorizes missingness into three distinct categories, namely: (1) Missing Completely at random (MCAR), (2) Missing at random (MAR), and (3) Not missing at random (NMAR) [10–12]. Handling incomplete data is a vital step in the analysis of medical datasets [13]. Among the various methods available, the simplest way to handle missingness is to delete records with the incomplete data and do the computation with complete records only. However, there are many drawbacks of this technique- it can lead to loss of information, it can affect the performance of classifiers, as deleted variables might be the deciding factor in predicting the disease, and, the collection of medical data involves time, money, and human efforts [14]. Imputation is an alternative approach employed to address missing data. This technique involves substituting missing values with estimated or imputed values. Imputation has been widely adopted as an efficient approach for managing incomplete data [15]. The task of imputing missing data holds significant importance across various domains, particularly in the medical or healthcare field. In this context, it becomes crucial to utilize all available data and avoid disregarding records solely due to the presence of missing values [16]. The most common method of imputation is filling in the missing data, with an average value of the missing variable, in all the observed cases of that variable. For numerical attributes, the mean value is utilized to replace missing values within the dataset, while for nominal attributes, the mode is used as a substitution approach. The advantage of this method is that the sample mean of that variable (missing variable) is not changed. However, the mean imputation technique is not suitable for multivariate analysis, as it underrepresents the variability in the data, and attenuates any correlations involving the imputed variable(s). ML)-based imputation techniques utilize the available variables to make predictions and estimate the missing data [14, 17]. These techniques employ the development of a predictive model for determining missing values in the datasets. ML-based models offer significant benefits, including heightened flexibility compared to traditional statistical models, enabling them

to capture intricate higher-order interactions within the data and consequently producing superior predictive outcomes.

The main contributions of this work include

- Comparison of three ML-based imputation techniques—KNNI, MICE, and MissForest.
- A comprehensive empirical analysis of three ML-based techniques-KNNI, MICE and MissForest, on UCI Diabetes Dataset for 10%-50% missing rate (MR).
- Performance analysis of the three imputation techniques is carried out on 16 datasets (one complete and fifteen imputed datasets), and evaluated using 11 evaluation criteria—accuracy, precision, recall, F1 score, Mcoff score, MAE, RMSE, R^2 values, Pearson correlation analysis, AIC, and BIC values.

The remainder of this paper is structured into five sections. Section 2 provides the background of ML based techniques. Section 3 provides a summary of the literature of different imputation techniques. In Section 4, the methodology employed in this study is presented and explained. The experimental setup, along with the results obtained and their analysis, is covered in Section 5. Impact of the work is covered in Section 6. In Section 7, we conclude by discussing our findings and future scope.

2 Background

This work explores KNN, MICE, and MissForest ML-based techniques. K Nearest Neighbour (KNN) is a ML- based imputation method. It computes the k-nearest neighbour for each of the missing values and imputes values from them. In numerical imputation, mean and weighted mean is used to replace the missing value while mode is used for binary or categorical variable. In weighted mean, greater weights are given to closer neighbours. The idea of KNN is that objects close to each other are potentially similar [13, 16]. Challenging issue is the selecting optimal value of k, and the other is selecting neighbours In KNN algorithm, generally Euclidean, Manhattan, Pearson etc. are used as similarity measure. Selection of similarity measure also plays a very important role in the overall performance of the algorithm [18]. The drawback of KNN is that it searches the whole databases to look for most similar instances. It is a limitation for large databases. Miss Forest is a machine learning-based imputation technique. It uses a Random Forest (RF) algorithm. It initializes the missing variables with mean or mode values. The variable under imputation is used as the target variable for building the RF model. The missing value is replaced by the prediction of the RF model. It is based on iterative approach, the process of looping through missing data points repeats several times [13]. Multiple Imputation by Chained Equations (MICE) is a prevalent method for executing multiple imputation because of its flexibility. In MICE, multivariate missing data are imputed on a attribute by attribute basis. called fully conditional specification (Van Buuren, 2007). This means that per variable imputations are created, such that for each incomplete variable a specified imputation model is required. In these imputation models, interactions can be modelled in two ways: first, by specifying models including interaction effects manually and second by imputing subgroups of the data separately. MICE consist of 3 steps, step1 is generation of multiple imputation, step2 is analyzing the imputed data and step3 is pooling the analysis results. Let us take a set of attributes, X1…… Xn, in which, some or all contains missing values. Initially, all missing

values are filled in at random. First attribute having missing value, In this example, X1 is regressed on the other attrbutes, X2,..., Xn. This is restricted to individuals with observed X1. The missing values in X1 are now replaced by simulated draws from the posterior predictive distribution of X1. This process is repeated for all other attributes X2…Xn. For attribute X2: X1, X3…Xn attributes will be considered. This cycle is repeated number of times, and creates one imputed dataset. The entire procedure is repeated m times, creating m imputed datasets. Each complete dataset is analyzed independently by MICE, then the results are pooled [19].

In MCAR, missing values are randomly distributed. KNNI can be effective when the missing values are irregularly related to other variables in the dataset. KNNI works on the assumption that the structure of the data remains similar for close instances. This makes it suitable for MCAR situations where missingness is not structured. MICE can handle MCAR beneficially as it imputes missing values built on observed values and the connections present in the dataset. MCAR presumes that missing values are not systematically related to any variables. Also, MICE is flexible in incorporating variable relationships. Hence MICE is suitable for handling MCAR missingness. Miss Forest ML method is powerful and can grasp complex relationships in the data. They work well even when missingness is random, as they can utilize information from other variables to envision missing values. The aggregate nature of Random Forests helps mitigate overfitting, making them suitable for imputation in datasets having a combination of MCAR and other missing data patterns.

The primary objective behind comparing machine learning-based imputation methods across four categories using 11 evaluation criteria in diabetes research is multifaceted, with key motivations including: a) The need to enhance data completeness and quality, b) The enhancement of predictive modeling for diabetes, c) The establishment of benchmark imputation methods tailored for diabetes research datasets, d) The utilization of standardized evaluation criteria to guarantee transparent and reproducible results when comparing imputation techniques, e) Empowering both clinicians and researchers with the requisite tools and knowledge to make well-informed decisions in the area of diabetes care and management.

3 Related work

In the existing literature, missing data in the medical field has been addressed through the application of statistical and machine learning-based imputation techniques. Statistical imputation assumes a normal distribution of the data and predicts missing values from the available data distribution. ML-based imputation techniques do not assume any specific data distribution and are capable of handling nonlinear relationships between variables [16, 20–24]. For instance, in a study on real breast cancer datasets, authors [16] employed statistical and ML-based methods to handle missing values. They utilized techniques like hot deck, mean, and hybrid imputation methods, as well as multilayer perceptron, K-nearest neighbor (KNN), and algorithms based on self-organizing map for handling missing data. In another study [25], researchers worked on medical datasets such as breast cancer, hepatitis, and diabetes datasets from the UCI repository. They proposed a novel hybrid prediction model that employed Simple K-means clustering to evaluate various imputation methods and select the superior one for filling in the missing data in the dataset. Similarly, in [26], the authors worked

with the hepatitis dataset, which contained an arbitrary pattern of missing values. They utilized principal component analysis and multiple imputation to fill in missing values having arbitrary pattern. Moreover, in [27], the authors also explored the hepatitis dataset and performed imputation using the bootstrap aggregating method. They compared the performance of this method with decision tree imputation, mode imputation, and mean imputation. The comparison demonstrated that the classifier yielded better results when using bootstrap aggregating imputation. Furthermore, in [28], researchers dealt with hepatitis and breast cancer datasets. They employed hot deck imputation for handling missing data and utilized an ensemble method for feature selection. By utilizing a neural network, the classification task was executed, resulting in an accuracy of 98.47% for the breast cancer dataset and 95.51% for the hepatitis dataset. Authors in [29] worked on a kidney dataset. They introduced the Weighted Average Ensemble Learning Imputation (WAELI) technique to fill in missingness and improve the disease prediction. RF classification and regression trees, and C4.5 were used to predict the missing values, and the resultant value was obtained by computing the weighted average of every model. In [30], a hybrid classifier was utilized for detecting retinal lesions caused by diabetic disease, where the dataset contained missing values. Another study [31] employed a novel hybrid classifier for predicting diabetes and employed multiple imputations to handle missing values. This hybrid classifier combined an adaptive model and logistic regression based on a fuzzy inference system. The deletion method is commonly used in the literature for handling missing values when predicting diabetes diseases. However, authors in [32] utilized Bayesian networks and TensorFlow factorization to process missingness in breast cancer datasets. They employed KNN, decision trees, and SVM for breast cancer recurrence prediction. Furthermore, researchers in [33] worked on the Iran diabetes dataset and proposed a hybrid imputation method based on single and multiple imputation. They compared the outcomes using three classifiers and evaluated the results based on accuracy, precision, recall, and F1 score. Lastly, in [34], a comparative study was conducted using decision trees, multilayer perceptron, KNN, and RF classifiers to enhance the accuracy of diabetes prediction. Mean imputation was employed for handling missingness. The precision of the imputation process in the healthcare domain can be further improved by incorporating domain expert knowledge [35]. The authors employed deep-learning techniques for predicting pneumonia [42]. Various machine learning-based imputation techniques are employed for medical datasets in [43–45].

One significant weakness in the literature is the limited discussion and comparison of specific machine learning algorithms and statistical methods used for imputation, which could be more detailed to enhance comprehensibility and applicability. Another critical issue is the lack of clear explanation and justification for the components and integration of the hybrid intelligent system, which hinders reproducibility. Finally, the complexity of the proposed model may limit its accessibility to researchers without specialized knowledge. Evaluating the performance of imputation methods can be challenging, as there is often no ground truth to compare against. This makes it difficult to assess the accuracy of imputed values. In the literature mostly model performance is selected as the evaluation criteria. Nevertheless, imputation fulfills a broader role within data analysis, and therefore, its effectiveness cannot be comprehensively assessed solely through model performance metrics. These metrics may fall short in encapsulating several critical aspects, including the extent of information loss, the introduction of bias, and the overall quality of imputed data (Table 1).

Table 1 Strengths and weaknesses of state-of-the-art

Paper	Imputation	Classifier	Weakness	Strength
16	Listwise deletion, mean, hot-deck, MI, MLP, Self-Organizing Map and KNNI	ANN	Imputation quality evaluation criteria is Prediction Accuracy	Worked on the data collected through the ''El A´lamo-I'' project
25	11 imputation methods including KNNI and SVMI	MLP	Only evaluation measures used is model performance, deletion is elected as the best imputation strategy for diabetes dataset	Presents a novel hybrid prediction model for diabetes, breast cancer and Hepatitis
26	Markov Chain Monte Carlo & Full Conditional Specification	MLP	Classification accuracy and error rate is used as evaluation criteria	Markov Chain Monte Carlo & Full Conditional Specification as Multiple imputation model
27	Mean, Mode, Decision tree, proposed bootstrap aggregation	KNN	Performance of imputation is evaluated using classification accuracy of KNN classifier	Work proposed a bootstrap aggregating based imputation approach for missing value handling
28	Hot deck imputation	backpropagation Neural network	Hot deck imputation method is used and model performance is considered as the evaluation criteria	Correlation based ensemble feature selection is used to select the relevant features from the dataset
29	Expectation Maximization, CART, Random Forest	Prediction model performance	Evaluation of imputation technique based on the proposed precision model performance	Weighted Average Ensemble Learning Imputation (WAELI) is proposed
31	Statistical Imputation	MLP	Evaluation is based on prediction model accuracy	proposed novel hybrid classifier for diabetes prediction
33	Markov Chain Monte Carlo, MICE and Expectation Maximization	MLP KNNCART	Evaluation of imputation technique based on prediction model performance	hybrid imputation model is proposed to impute missing data
34	Mean imputation is used	DT, KNN, RF, MLP	Evaluation is done using prediction accuracy	Comparative study of classifiers and feature selection methods
35	Expectation Maximization Imputation	Regression methods along with Ewing formula are employed to categorize the classes of CAN	Evaluation criteria is prediction accuracy	Regression methods are employed to categorize the classes of (CAN)

Table 1 (continued)

Paper	Imputation	Classifier	Weakness	Strength
Our paper	KNNI, MICE MissForest	RF, SVM, AdaBoost, XGB	Handled MCAR mechanism only	Imputation performance is evaluated using four criteria i.e. 1) model performance, 2) imputation error rate MAE, RMSE, R^2 values, 3) Pearson correlation analysis and, 4) model selection basis BIC, AIC values). Model performance includes accuracy, precision, recall, F1 score, and Mcoff score of four ML classifiers. i.e. total 11 criteria are used for the evaluation

4 Methodology

In this study, we used Pima Indians Diabetes Dataset, which is sourced from the UCI containing 768 records. However, 376 records had missing values in one or more variables, so we deleted those records, and processing was done using 392 complete records. Thereafter, we generated synthetic missingness in this dataset, using the MCAR mechanism, to generate five incomplete datasets having 10%-50% MR. This missingness was generated in multivariate configuration, in more than one variable, using the binomial distribution. We used three ML-based imputation techniques- KNNI, MICE, and MissForest, to impute the missing data in five incomplete datasets, to generate fifteen imputed datasets – five datasets of KNNI imputation, five datasets of MICE imputation and five datasets of MissForest imputation. The design process for comparison of imputation techniques is shown in Fig. 1.

We evaluated the performance of KNNI, MICE, and MissForest in four categories- 1) Diabetes Prediction Model Performance, 2) Imputation error rate, 3) Correlation analysis, 4) Model selection basis.

1. Diabetes prediction model performance: This model was built with one complete dataset and fifteen imputed datasets using four classifiers- RF, SVM, AdaBoost, and XGBoost (XGB). These four classifiers are widely used for machine learning imputation techniques on medical datasets. RF is an ensemble method, SVM is a linear and non-linear classifier, AdaBoost is an ensemble boosting method, and XGB is a gradient boosting algorithm. This diversity helps assess how different types of classifiers react to imputed data. The prediction performance of four classifiers with imputed datasets is compared with one complete dataset, using five evaluation metrics- accuracy, precision, recall, F1 score, and Mcoff score.
2. Imputation error rate: We evaluated the quality of imputation of KNNI, MICE & MissForest techniques using metrics-MAE, RMSE, and R^2, by comparing one complete dataset and fifteen imputed dataset values. MAE, RMSE, and R^2 values are calculated for KNNI, MICE & MissForest techniques for 10% to 50% MR.
3. Correlation analysis: It is performed to identify the imputation technique suitable to grasp the intricate connection among various variables in the diabetes dataset, and

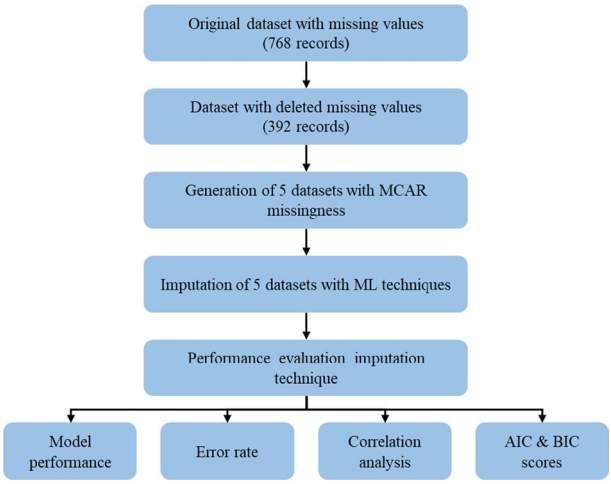

Fig. 1 The experimental design process for comparison of imputation techniques

produce more accurate results. The Pearson correlation coefficient of all the variables in the fifteen imputed datasets is calculated, and compared with the Pearson correlation coefficient values of all the variables, in one complete dataset.

4. Model selection basis: We selected the best model after calculating & comparing the AIC & BIC scores of the full model and step model of one complete and fifteen imputed datasets. The full model is constructed with all the variables, & step model is constructed using stepwise regression, which selects a subset of variables to improve the performance of the model, and build the step model.

5 Experimental setup and results

The objective of this experiment was to conduct a comparative analysis of MCAR (Missing Completely at Random) Multivariate Missing patterns and assess the effectiveness of three machine learning-based imputation techniques in addressing them for 10%—50% MR. This study evaluates the performance of KNNI, MICE, and MissForest using four categories- Diabetes Prediction Model Performance, Imputation error rate, Pearson Correlation analysis, and Model selection based on AIC and BIC scores. The Diabetes Prediction Model performance of KNNI, MICE, and MissForest is evaluated with four ML classifiers namely RF, SVM, AdaBoost & XGB. Diabetes prediction is carried out for one complete dataset and fifteen imputed datasets by RF, SVM, AdaBoost, and SVM classifiers. The performance of the imputation techniques with four classifiers is evaluated using five evaluation criteria- accuracy, precision, recall, relative F1 score, and Mcoff score. Imputation error is evaluated using MAE and RMSE and R^2 values, Pearson correlation analysis of the variables of one complete dataset and fifteen imputed datasets is calculated, and compared to check the preservation of the relationship between variables, before & after imputation. Model selection is based on AIC and BIC scores. The experiments conducted in this study utilized the Pima Indians Diabetes dataset, obtained from the UCI repository [36]. This dataset consists of a total of 768 patient records, all of which are female. Among these records, there are 268 cases of diabetic patients and 500 cases of non-diabetic patients. The dataset provided in Table 2 consists of information on eight attributes, including glucose, blood pressure, skin thickness, insulin, and BMI. To handle missing values, records with missing entries were removed, resulting in a dataset containing 392 records that were processed for further analysis. Out of 392 patient records used for analysis, 130 records belong

Table 2 Overview Pima dataset attributes and missingness

Attribute	Mean of attribute	Missing values	Missing %	Correlation value with output class
Pregnancies	3.85	0	0	0.222
Glucose	120.89	5	1	0.467
Blood Pressure	69.11	35	5	0.065
Skin Thickness	20.54	227	30	0.075
Insulin	79.80	374	49	0.131
BMI (Body Mass Index)	31.99	11	1	0.293
Diabetes Pedigree Function	0.47	0	0	0.174
Age	33.25	0	0	0.238

to diabetes present cases and 262 records belong to diabetes absent cases. In the dataset containing 392 complete records, missingness was artificially generated. The experimentations were accomplished using Python 3.8 on the Anaconda Jupyter Notebook platform.

To create five incomplete datasets, various missing rates ranging from 10 to 50% were artificially introduced into the input variables. It's important to note that the output variable remained intact and was not affected by the missing values. The next step involved imputing the missing values in these five datasets using the KNNI, MICE, and MF techniques. As a result, there were five imputed datasets for KNNI and MICE and MF imputations, amounting to a total of fifteen imputed datasets. Additionally, one complete dataset without any missing values was included, resulting in a total of sixteen datasets for experimentation. The sample size used for these experiments was 392. In these 392 samples of the Pima Indians Diabetes dataset, the generated artificial missingness was produced through the MCAR mechanism, while the true data exhibits characteristics that fall in between MAR and MCAR [7]. This missingness was generated randomly, in a multivariate pattern, which means missingness is present in more than one variable of the dataset, with the binomial distribution.

Results To evaluate the comparative effectiveness of KNNI, MICE, and MissForest imputation techniques, our experimental design was formulated. In this study, a comparative analysis was conducted to assess the performance of three imputation methods across four categories: Model Performance, imputation error rate (MAE, RMSE, R^2 values), Pearson correlation analysis, and Model Selection based on AIC and BIC values. To evaluate the model performance, a ten-fold cross-validation technique was employed. The entire dataset was subjected to 10 repetitions of the experiment, with each sample being tested. The average of the outcomes from all 10 iterations was then chosen as the ultimate result. The performance of the Diabetes Prediction Model was evaluated by considering metrics such as accuracy, precision, recall, F1 score, and Mcoff score with RF, SVM, AdaBoost & XGB classifiers (4.1–4.5), Imputation error rate, Coefficient of Determination of complete datasets and imputed datasets are compared (4.6), Correlation analysis (4.7), Model selection based on AIC & BIC values for various missing rate is carried out (4.8).

5.1 Relative performance analysis of prediction accuracy

Prediction accuracy is obtained by dividing the number of correct predictions by the size of the dataset.

$$Accuracy = (TP + TN) \div (TP + TN + FP + FN)$$

TP, TN, represents True Positive & True Negative and FP and FN represents False Positive & False Negative values respectively

$$Relative\ Accuracy = 100 \times ((AO - (AO - AM))) \div AO$$

In the context where all features are available and known, AO represents the prediction accuracy of the complete dataset, The prediction accuracy, denoted by AM, was measured after applying each respective imputation method to fill in missing data. As depicted in Fig. 2, it is apparent that the MissForest algorithm outperforms other imputation techniques in four out of the five cases i.e. in 80% of cases with varying percentages of missing data.

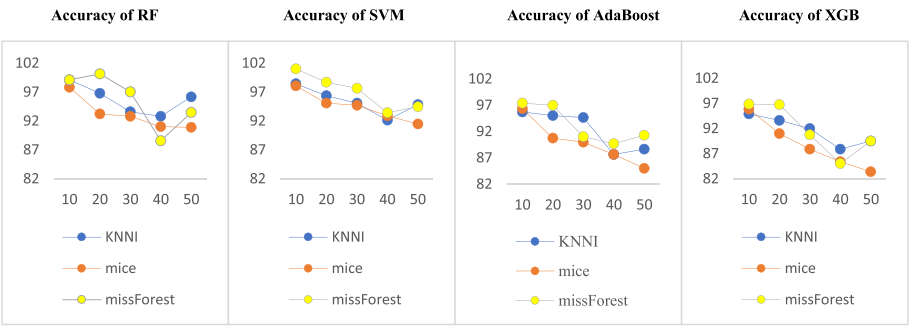

Fig. 2 Comparing relative differences of prediction accuracy between Original and Imputed dataset for four classifiers

5.2 Relative performance analysis of precision

Precision is a measure of how many of the positive predictions made are correct i.e. TP. TP + FP is the Number of patient models predicted with diabetes

$$Precision = TP \div (TP + FP)$$

$$Relative\ Precision\ Score = 100 \times ((PO - (PO - PM)) \div PO)$$

In the context where all features are available and known, PO denotes the precision score of the complete dataset, while PM represents the precision score measured after applying each respective imputation method to fill in missing data. From Fig. 3, it is evident that the KNNI algorithm performs better than other imputation techniques, in three out of five missing % cases i.e., 60% of cases, and MissForest performance is better in 40% of cases.

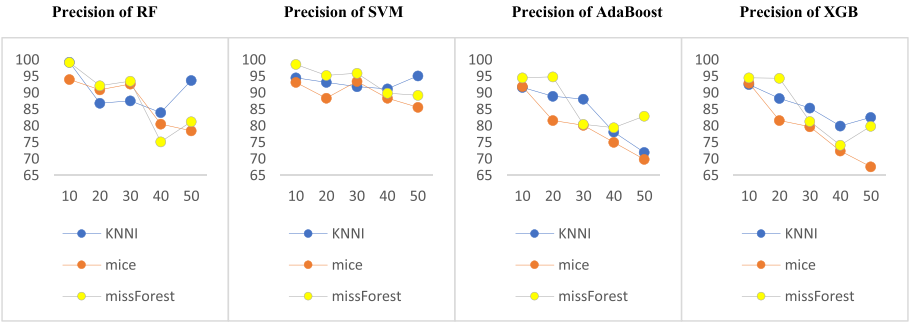

Fig. 3 Comparing relative differences of precision score between Original and Imputed datasets for four classifiers

5.3 Relative performance analysis of recall

The recall is a measure of how many positive cases the classifier has accurately predicted, It is very important in medical domains as we want to minimize the chance of missing positive cases.

$$Recall = TP \div (TP + FN)$$

where TP (True Positive) is no. of correctly predicted patient with diabetes and TP + FN (false Negative) is total no. of patients with diabetes in the dataset.

$$Relative\ Difference\ Recall\ Score = 100 \times ((RO - (RO - RM)) \div RO)$$

In the context where all features are available and known, RO refers to the recall score of the complete dataset and RM represents the recall score after applying the corresponding imputation method to impute missing values. From Fig. 4, it is evident that the MissForest imputation technique outperforms other imputation techniques in three out of five missing % of cases i.e., 60% of cases. It is also observed that the MissForest imputation technique gives the best performance with the SVM classifier.

5.4 Relative performance analysis of F1 score

The F1 score, a comprehensive evaluation metric, accounts for both precision and recall, making it suitable for imbalanced datasets where precision and recall must both be considered.

$$F1\ Score = 2 \times ((precision \times recall)/(precision + recall))$$

$$Relative\ Difference\ F1\ Score = 100 \times ((FO - (FO - FM)) \div FO)$$

In the context where all features are available and known, FO refers to the F1 score of the complete dataset and FM is the F1 score after applying the corresponding imputation method to impute missing values. From Fig. 5 it is evident that the MissForest algorithm performs better than other imputation techniques in 100% of cases. It is also observed that the MissForest imputation algorithm gives the best performance with the SVM classifier.

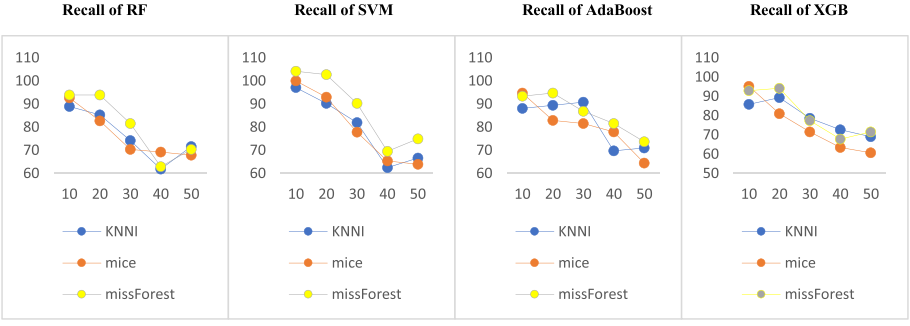

Fig. 4 Comparing relative differences of recall scores between Original and Imputed datasets for four classifiers

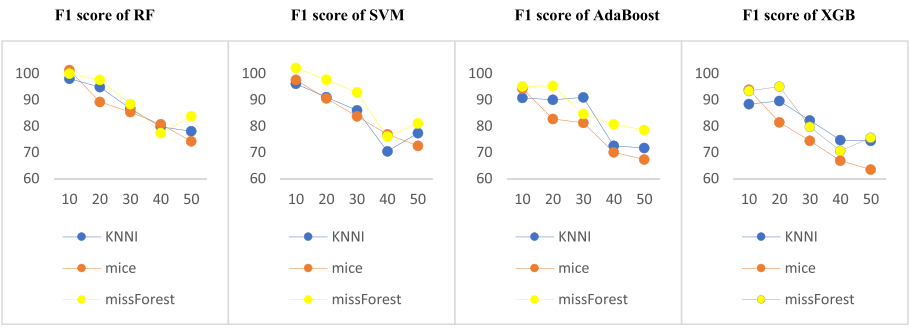

Fig. 5 Comparing relative differences of F1 score between Original and Imputed datasets for four Classifiers

5.5 Relative performance analysis of Mcoff-score

Mcoff is considered a consistent evaluation metric since it yields a high score only when the prediction exhibits excellent performance across all four categories of the confusion matrix.

$$Relative\ Mcoff\ Score = ((MO - (MO - MM)) \div MO)$$

In the context where all features are available and known MO refers the Mcoff score of the complete dataset, and MM is the Mcoff score after applying the corresponding imputation method to impute missing values. From Fig. 6, it is evident that among the other imputation techniques, the MissForest algorithm shows better performance in four out of five cases i.e., 80% of cases. It is also observed that the MissForest imputation algorithm gives the superior results with the SVM classifier. The overall performance of imputation techniques and classifiers is shown in Figs. 7 and 8 respectively. After comparison of model performance across missing rates ranging from 10 to 50% [Figs. 2, 3, 4, 5, and 6], it becomes evident that the performance of four classifiers, as evaluated using five criteria, is better at the 50% MR compared to the 40% MR. Generally, classifier performance tends to decline as the missing rate increases. This occurrence could be attributed to the synthetic generation of MCAR-type missingness, where important predictive features might exhibit a higher degree of missingness in the case of the 40% MR.

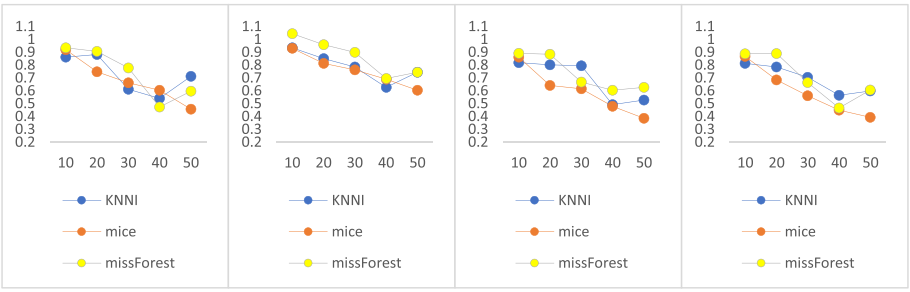

Fig. 6 Comparing relative differences in Mcoff scores between original and imputed datasets for four classifiers

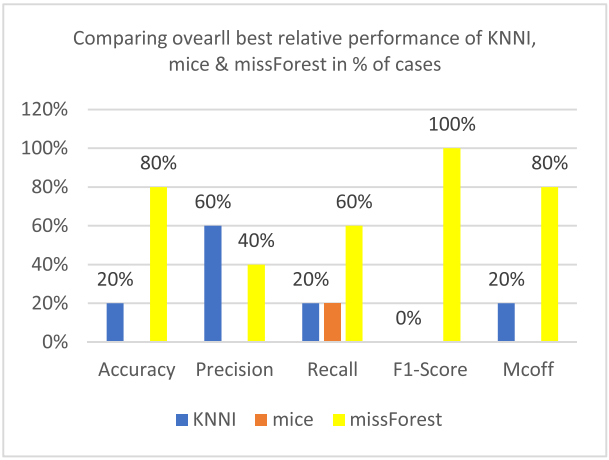

Fig. 7 Comparing overall performance of Imputation techniques

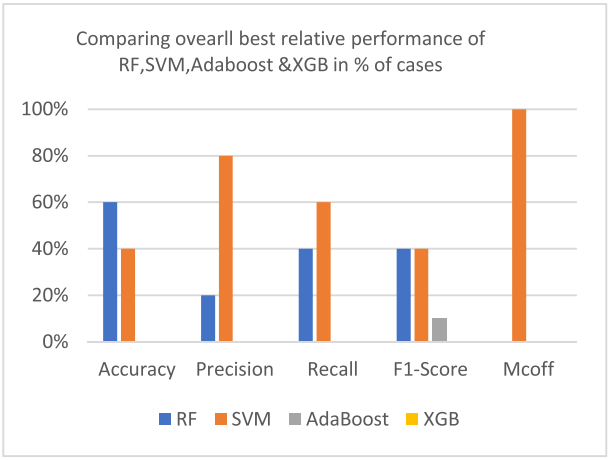

Fig. 8 Comparing overall performance of Classifiers

5.6 Relative performance analysis of imputation techniques by MAE, RMSE, R^2 values

The accuracy of the imputation method is evaluated using various metrics that assess the discrepancy between the imputed values and the actual values of missing data. One commonly used metric is the MAE, which calculates the average absolute difference between the imputed values and the true values. A lower MAE value indicates better performance. Another frequently employed metric is the RMSE, which measures the square root of the average of the squared differences between the imputed values and the true values. A lower RMSE value indicates better performance in capturing the differences between the model-predicted values and the observed values. Additionally, R^2 is utilized to measure the proportion of the variance in the true values that can be explained by the imputed values. A higher R^2 value signifies a stronger correlation and a better representation of the true values by the imputed values. A higher value of R^2 indicates better performance. Complete dataset values are compared with fifteen imputed datasets of 10%-50% MR filled with KNNI, MICE, and MissForest techniques to calculate MAE, RMSE, and R^2 values.

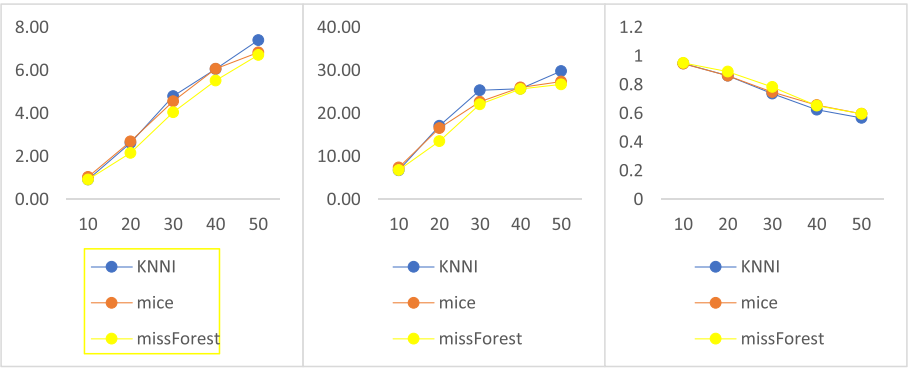

Fig. 9 Comparison of MAE, RMSE and R^2 values for various Imputation techniques

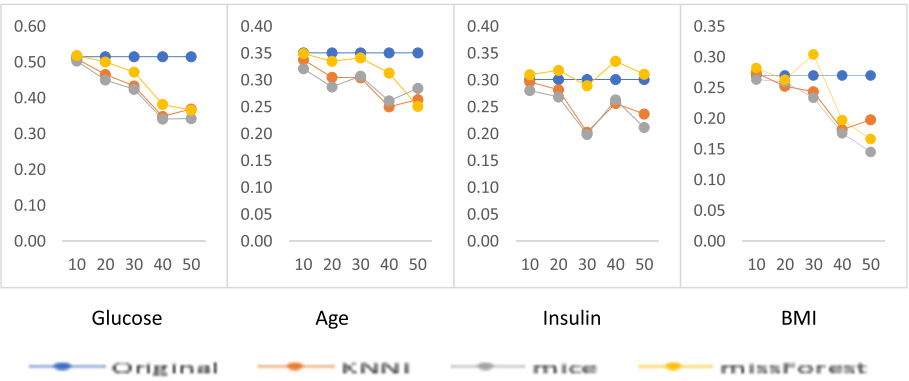

Fig. 10 Comparison of Pearson correlation coefficient Values of Glucose, Age, Insulin & BMI for various Imputation techniques

It is observed that the MissForest Imputation method achieved lower MAE, and RMSE in 100% of MR cases and higher R^2 in 100% of MR cases, as compared to KNNI and MICE. This revealed that the performance of MissForest is better than the other two imputation techniques, in all of these three evaluation criteria as shown in Fig. 9.

5.7 Relative performance analysis of imputation techniques by correlation analysis

Correlation analysis quantifies the association between the imputed values and other variables within the dataset. We evaluated the performance of KNNI, MICE, and MissForest imputation techniques by comparing Pearson correlation coefficient values of Glucose, Age, Insulin, BMI, Pregnancies, skin thickness, Diabetes Pedigree Function, and BP variables of the complete dataset and fifteen imputed datasets, to check if the imputed values are correlated with other variables in the dataset. Correlation analysis of three imputation techniques shows that the MissForest imputation technique can capture the complex relationship between the variables, like the complete dataset, for all the variables, as compared to MICE and KNNI imputation techniques. Results are shown in Figs. 10 and 11.

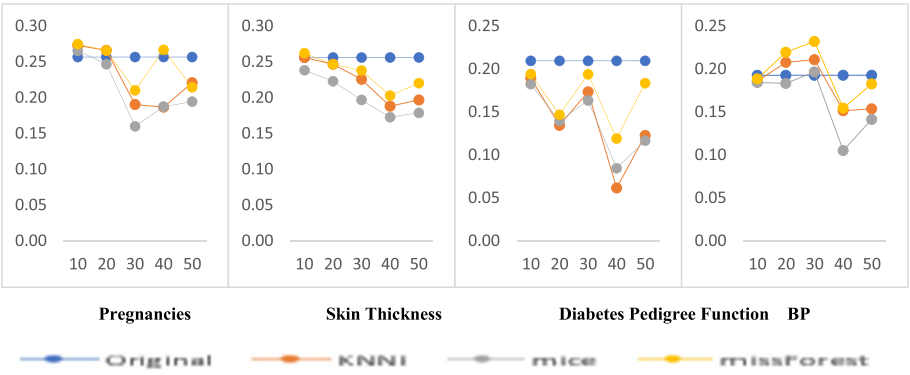

Fig. 11 Comparison of Pearson correlation coefficient Values of Pregnancies, Skin Thickness, Diabetes Pedigree Function & BP for various Imputation techniques

5.8 Relative performance analysis of imputation techniques by AIC and BIC scores

5.8.1 AIC

It is a model selection principle proposed by Akaike in 1973. AIC helps in selecting a model, by estimating the quality of each model given as an input to it. AIC evaluates the effectiveness of a model based on the extent to which it preserves information, with higher quality models retaining less lost information. AIC accounts for the potential risks of overfitting and underfitting in model estimation, with lower AIC values indicating a more optimal model fit. AIC penalizes complex models less, so less score is given to the complex model, and finally, a complex model is selected. The full model is the model with all the variables of the dataset [37–39]. The step model is constructed using stepwise regression which selects a subset of variables and builds the step model which gives the best performing model by iteratively adding. and deleting variables. Results are shown in Figs. 12 and 13.

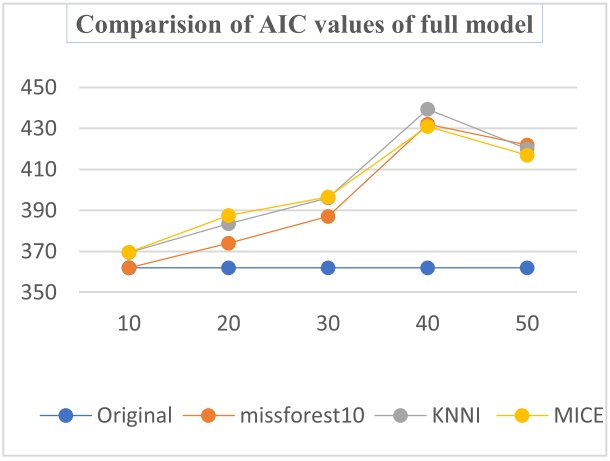

Fig. 12 Comparison of AIC values of full model

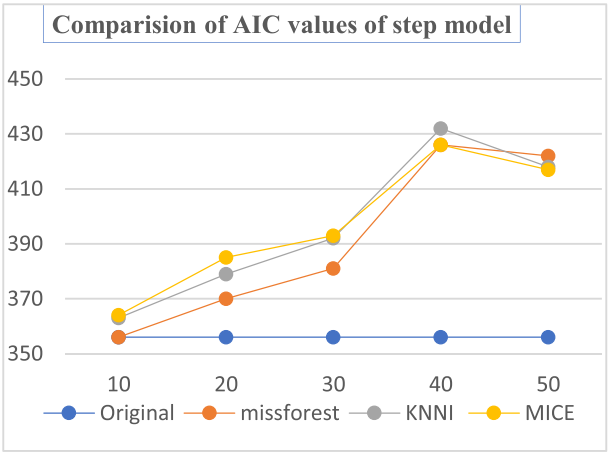

Fig. 13 Comparison of AIC values of step model

5.8.2 BIC

Schwarz proposed the Bayesian Information Criterion (BIC) in 1978 as a model selection principle, which serves as an asymptotic approximation to a transformed Bayesian posterior probability of a candidate model.

$$AIC = -2ln(maximum\ likelihood) + 2m$$

$$BIC = -2ln(maximum\ likelihood) + mln(n)$$

The best model is selected based on the minimum value of AIC or BIC, where AIC and BIC are estimated using the number of estimated parameters (m) and the number of observations (n). BIC penalizes the model more as compared to AIC for its complexity, BIC selects the less complex one [40, 41].

The full model and stepwise regression model are constructed for the complete dataset and fifteen imputed datasets for 10–50% MR. AIC and BIC score comparison of the full model and step model is carried out for KNNI, MICE, and MissForest. AIC and BIC score analysis show that the performance of MissForest is better than MICE and KNNI imputation techniques, for the full and step model. Results are shown in Figs. 14 and 15.

6 Impact of the work

Diabetes is a chronic disease that requires continuous monitoring and management. Departments in large hospitals monitoring chronic diseases generate a lot of data with probability of missingness. Addressing missingness in a scientific manner helps in reducing knowledge loss and accurate decision making in various healthcare domains. This work has assessed imputation techniques using 11 evaluation criteria which provided a holistic understanding of imputation techniques' performance. Conducting Pearson correlation analysis allows to understand the relationships between variables in the dataset before and after the imputation. Other 10 evaluation criteria also capture different aspects of imputation offering a comprehensive view of strengths and weaknesses which can help identifying the best imputation technique and classifier which can ensure that datasets are more complete and

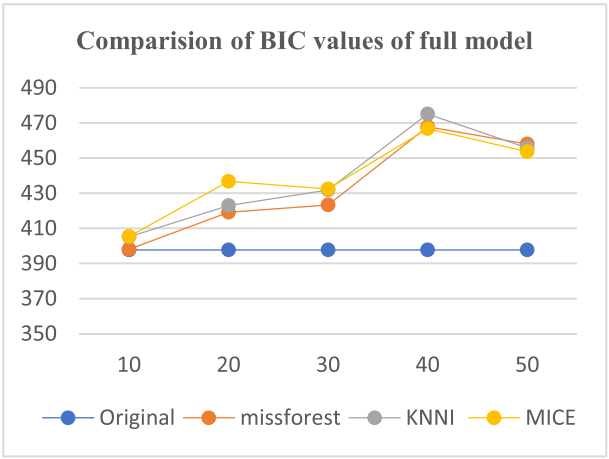

Fig. 14 Comparison of BIC values of full model

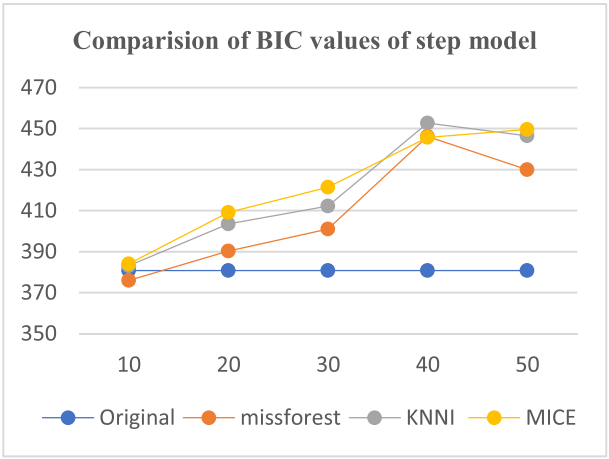

Fig. 15 Comparison of BIC values of step model

of higher quality which is crucial for accurate analysis, and also help in improving disease prediction models and accurate decision making.

7 Conclusion and future scope

In this work, a comparative analysis of three ML-based imputation methods was performed on the Pima Indian dataset. Experimental evidence confirmed that the MissForest imputation technique performed better in eleven evaluation criteria, as compared to the other imputation techniques. It was also found that the SVM classifier performed better than RF, XGB, and AdaBoost classifiers in the precision, recall, F1 score, and Mcoff. The empirical analysis for all five MR (10% to 50%) cases, using MissForest, KNNI, and MICE techniques, revealed that the MissForest method performed better in accuracy & Mcoff in 80% of cases, better in precision & recall in 60% of cases, better in F1 score, MAE, RMSE, R^2, AIC, BIC values in 100% of cases. The Pearson Correlation Coefficient analysis of the input variables also revealed that MissForest techniques were able to capture the complex

relationship between all the variables in the diabetes dataset. Overall, our empirical evidence confirms that MissForest is a better ML-based imputation technique, for handling missing data in diabetes datasets. The use of accurate imputation techniques can improve the quality of diabetes research, by ensuring that missing data does not compromise the validity of research results. In this work, we exclusively addressed the MCAR missing mechanism. However, we intend to address this limitation in the future by incorporating methods to handle the MAR mechanism as well as introducing an ensemble imputation approach and explore other ML based methods that is capable of effectively managing both MCAR and MAR missingness in different diseases. Also, future direction of this study involves enhancing imputation by integrating medical expertise and developing real-time imputation applications for missing data, in clinical settings where prompt decision-making is essential.

Data availability The data used in this article will be shared on request made to the corresponding author.

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, Stein C et al (2022) IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract 183:109119
2. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW (2019) Global trends in diabetes complications: a review of current evidence. Diabetologia 62:3–16
3. Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, Gehlot A, Rashid M, Alshamrani SS, AlGhamdi AS (2022) An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment. Appl Sci 12(8):3989
4. Fasihi M, Nadimi-Shahraki MH (2020) Multi-class cardiovascular diseases diagnosis from electrocardiogram signals using 1-D convolution neural network. In: 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE
5. Zamani H, Nadimi-Shahraki MH (2016) Swarm intelligence approach for breast cancer diagnosis. Int J Comput Appl 151(1):40–44
6. Nadimi-Shahraki MH et al (2021) B-MFO: a binary moth-flame optimization for feature selection from medical datasets. Computers 10(11):136
7. Ramli MNN et al (2013) Roles of imputation methods for filling the missing values: A review. Adv Environ Biol 7(12 S2):3861–3870
8. Cios KJ, William Moore G (2002) Uniqueness of medical data mining. Artif Intell Med 26(1–2):1–24. https://doi.org/10.1016/s0933-3657(02)00049-0
9. Newman CBD (1998) UCI repository of machine learning databases. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html
10 Liu Y, Brown SD (2013) Comparison of five iterative imputation methods for multivariate classification. Chemom Intell Lab Syst 120:106–115
11. Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
12. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR (2018) Characterizing and managing missing structured data in electronic health records: data analysis. JMIR Med Inform 6(1):e11. https://doi.org/10.2196/medinform.8960
13. Lin WC, Tsai CF (2020) Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 53:1487–1509. https://doi.org/10.1007/s10462-019-09709-4
14. Zhang S (2011) Shell-neighbor method and its application in missing data imputation. Appl Intell 35:123–133. https://doi.org/10.1007/s10489-009-0207-6

15. Thomas RM, Bruin W, Zhutovsky P, van Wingen G (2020) Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders. In Machine learning. Academic Press, pp 249–266
16. Jerez JM et al (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 50(2):105–15
17. Zhang h (2012) Nearest neighbor selection for iteratively kNN imputation. J Syst Softw 85(11):2541–2552. ISSN 0164–1212
18. Zeng, Xie D, Liu R, Li X (2017) Missing value imputation methods for TCM medical data and its effect in the classifier accuracy. 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). pp 1–4. https://doi.org/10.1109/HealthCom.2017.8210844
19. Doove LL et al (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal 72:92–104
20. Little RJ, Rubin DB (2019) Statistical analysis with missing data. John Wiley & Sons
21. Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8(1):3–15
22. Buuren SV (2018) Flexible imputation of missing data. CRC Press
23. Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20(1):40–49
24. Kim J, Kim H (2018) Comparison of statistical and machine learning methods for imputing missing data in electrical impedance tomography. Comput Biol Med 92:8–15
25. Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2015.02.050
26. Ramezani R, Maadi M, Khatami SM (2018) A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alex Eng J 57(3):1883–1891. ISSN 1110–0168
27. Radhakrishnan S, Priyaa DS (2015) An ensemble approach on missing value handling in hepatitis disease dataset. Int J Comp Appl 130:23–27. https://doi.org/10.5120/ijca2015907197. Sridevi Radhakrishnan and Shanmuga D Priyaa. Article: An Ensemble approach on Missing Value Handling in Hepatitis Disease Dataset. International Journal of Computer Applications 130(17):23–27, November 2015. Published by Foundation of Computer Science (FCS), NY, USA
28. Elgin Christo VR, Khanna Nehemiah H, Minu B, Kannan A (2019) Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network. Comput Math Methods Med 2019:7398307. https://doi.org/10.1155/2019/7398307
29. Arasu SD, Thirumalaiselvi R (2017) A novel imputation method for effective prediction of coronary Kidney disease. In: 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), Chennai, India. pp 127–136. https://doi.org/10.1109/ICCCT2.2017.7972256
30 UsmanAkram M, Khalid S, Tariq A, Khan SA, Azam F (2014) Detection and classification of retinal lesions for grading of diabetic retinopathy. Comp Biol Med 45:161–171. ISSN 0010-4825
31. Ramezani R, Maadi M, Khatami SM (2018) A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alex Eng J 57(3):1883–1891
32. Vazifehdan M, Moattar MH, Jalali M (2019) A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. J King Saud Univ - Comp Inf Sci 31(2):175–184. ISSN 1319-1578
33. Nadimi-Shahraki MH, Mohammadi S, Zamani H, Gandomi M, Gandomi AH (2021) A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis. Electronics 10(24):3167
34. Saxena R, Sharma SK, Gupta M, Sampada GC (2022) A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Comput Intell Neurosci 2022:3820360. https://doi.org/10.1155/2022/3820360
35. Abawajy J et al (2013) Predicting cardiac autonomic neuropathy category for diabetic data with missing values. Comp Biol Med 43(10):1328–33. https://doi.org/10.1016/j.compbiomed.2013.07.002
36 Liew AW-C et al (2011) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform 12(5):498–513. https://doi.org/10.1093/bib/bbq080
37. Learning UM (2016) Pima indians diabetes database. kaggle. com/uciml/pima-indians-diabetes-database
38. Rekabdar B, Albright DL, McDaniel JT, Talafha S, Jeong H (2022) From machine learning to deep learning: a comprehensive study of alcohol and drug use disorder. Healthcare Analytics 2:100104
39. An H, Gu L (1989) fast stepwise procedure of selection of variables by using AIC & BIC criteria. Acta Math Appl Sin 5(1):60–67
40. Yamashita T, Yamashita K, Kamimura R (2007) A stepwise AIC method for variable selection in linear regression. Comm Statist Theory Methods 36(13):2395–2403

41. Chakrabarti A, Ghosh JK (2011) AIC, BIC and recent advances in model selection. Philos Stat 583–605
42. Swetha KR, Niranjanamurthy M, Amulya MP, Manu YM (2021) Prediction of pneumonia using big data, deep learning and machine learning techniques. In: 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE 1697–1700. https://doi.org/10.1109/ICCES51350.2021.9489188
43. Reshma VK, Khan IR, Niranjanamurthy M, Aggarwal PK, Hemalatha S, Almuzaini KK, Amoatey ET (2022) Hybrid block-based lightweight machine learning-based predictive models for quality preserving in the Internet of Things- (IoT-) based medical images with diagnostic applications. Comput Intell Neurosci 2022:Article ID 8173372, 14 pages. https://doi.org/10.1155/2022/8173372
44. AkkemY, BiswasSK, Varanasi A (2023) Smart farming using artificial intelligence: A review. Eng Appl Artif Intell 120:105899. ISSN 0952–1976. https://doi.org/10.1016/j.engappai.2023.105899
45. Akkem Y, Biswas SK, Varanasi A (2023) Smart farming monitoring using ML and MLOps. In: Hassanien AE, Castillo O, Anand S, Jaiswal A (eds) International Conference on Innovative Computing and Communications. ICICC 2023. Lecture Notes in Networks and Systems, vol 703. Springer, Singapore. https://doi.org/10.1007/978-981-99-3315-0_51