Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 06 June-2026, Page No.- 10666- 10669

DOI: 10.47191/etj/v11i06.43, I.F. – 8.595

Predicting Missing Values in Randomly Incomplete Data Using Artificial Intelligence Algorithms: An Experimental Study with More Than 95% Normalized Accuracy

Ahmed K Alhebshi

Department of Engineering and Computing, University of Science and Technology Hadramout, Yemen

ABSTRACT: Missing values often appear in operational datasets long before any statistical analysis or machine learning model is built. When the incomplete entries are treated with overly simple rules, such as row deletion or column means, the resulting dataset may no longer reflect the original relationships among variables. This paper presents an experimental framework for estimating randomly missing numerical values using artificial intelligence and machine learning techniques. The Wisconsin Diagnostic Breast Cancer dataset was used as a reproducible benchmark. Three random missingness levels were simulated: 10%, 20%, and 30%. Mean imputation, K-Nearest Neighbors imputation, and Bayesian Ridge-based Multivariate Imputation by Chained Equations were evaluated using Mean Absolute Error, Root Mean Squared Error, and Normalized MAE Accuracy. The Bayesian Ridge MICE approach produced the strongest results, achieving 97.51%, 96.93%, and 96.60% Normalized MAE Accuracy at 10%, 20%, and 30% missingness, respectively. These findings show that a multivariate AI-based imputation workflow can recover randomly missing values with more than 95% normalized accuracy in this experimental setting and can outperform the tested baseline methods.

KEYWORDS: missing values; artificial intelligence; machine learning; data imputation; MICE; Bayesian Ridge; KNN; random missing data.

1. INTRODUCTION

Incomplete data is a practical problem in many information systems. Records may be partially lost because of typing mistakes, sensor interruptions, delayed reporting, survey non-response, or software integration issues. Even when the amount of missing data seems small, the effect on later analysis can be substantial because many algorithms assume that all input features are available.

A common response is to remove incomplete observations or replace missing entries with simple summaries such as the mean or median. These approaches are convenient, but they do not use the relationships that exist among variables. As a result, they may flatten the distribution of the data and hide important patterns. More advanced imputation methods attempt to estimate missing values by learning from the observed part of the dataset.

Artificial intelligence algorithms are suitable for this task because they can model multivariate structure. Instead of treating each column as isolated, these algorithms use other available features to infer the missing value. This study examines that idea through a controlled experiment in which complete data are first normalized, values are then removed at random, and several imputation methods are compared against the original hidden values.

2. PROBLEM STATEMENT

The central problem addressed in this paper is the reliable reconstruction of missing numerical values when the missingness is random. The research question is: To what extent can AI-based imputation methods predict randomly missing values more accurately than conventional baseline methods while preserving the numerical structure of the dataset?

3. RESEARCH OBJECTIVES

- To develop a reproducible experimental workflow for random missing value prediction.
- To compare a simple statistical baseline with machine learning-based imputation methods.
- To measure imputation error using MAE and RMSE on the actual masked values.
- To report Normalized MAE Accuracy after min-max scaling of all numerical features.
- To identify the strongest method within the tested experimental setup.

4. LITERATURE REVIEW

Research on missing data distinguishes between different missingness mechanisms, including Missing Completely at Random, Missing at Random, and Missing Not at Random.

The present paper focuses on the first case because the incomplete entries are generated through a controlled random masking process. This design allows the original values to remain known for evaluation while still creating realistic incomplete input data.

Simple imputation methods, including mean and median replacement, are often used as first baselines. Their advantage is speed and interpretability, but their weakness is that they do not learn from other variables. Similarity-based techniques such as K-Nearest Neighbors improve on this idea by estimating missing entries from nearby records. However, KNN performance can be affected by feature scaling, dimensionality, and the availability of sufficiently similar observations.

Multivariate Imputation by Chained Equations provides a more flexible strategy. It treats variables with missing values as prediction targets and repeatedly estimates them from the remaining variables. This iterative modeling approach can preserve inter-feature relationships more effectively than one-column replacement. Tree-based methods and neural network-based methods have also been investigated in the literature, including missForest and DataWig, showing the continued importance of machine learning for data completion tasks.

5. DATASET DESCRIPTION

The experiment used the Wisconsin Diagnostic Breast Cancer dataset, a widely used benchmark dataset available through the UCI Machine Learning Repository and the scikit-learn library. It contains 569 observations and 30 continuous numerical features computed from digitized images of breast mass cell nuclei. The class label was retained only as an observed auxiliary covariate during imputation; it was not included in the error calculation.

Table 1. Dataset description.

Property	Description
Dataset name	Wisconsin Diagnostic Breast Cancer
Source	UCI Machine Learning Repository / scikit-learn
Number of instances	569
Number of numerical features	30
Missingness mechanism in this study	Artificial Missing Completely at Random
Missingness rates	10%, 20%, and 30%

6. METHODOLOGY

6.1 Data Normalization

All numerical features were scaled to the interval [0, 1] using min-max normalization. This preprocessing step ensured that error values were comparable across features with different original units and made it possible to define a clear normalized accuracy measure.

6.2 Random Missing Value Generation

For each missingness level, a random binary mask was generated over the feature matrix. Masked entries were replaced with missing values, while the original values were stored separately for evaluation. Five random trials were used for each missingness level in order to reduce the effect of a single random split.

6.3 Imputation Algorithms

Three methods were evaluated. Mean imputation was used as the statistical baseline. KNN imputation represented a similarity-based machine learning approach. Bayesian Ridge MICE represented the multivariate iterative AI-based approach because it predicts each incomplete variable from the remaining available variables and repeats the process until stable estimates are produced.

Table 2. Experimental imputation configuration.

Method	Main parameters	Role in the experiment
Mean Imputation	strategy = mean	Baseline statistical method
KNN Imputation	n_neighbors = 5; weights = distance	Similarity-based imputation
Bayesian Ridge MICE	max_iter = 20; estimator = Bayesian Ridge	Iterative multivariate AI-based imputation

6.4 Evaluation Metrics

The predicted entries were compared only with the original values that had been masked. Mean Absolute Error was used to measure the average magnitude of prediction error, while Root Mean Squared Error was used to penalize larger errors. Because the feature values were normalized to [0, 1], Normalized MAE Accuracy was computed as follows:

Normalized MAE Accuracy = (1 - MAE) × 100

This metric should be interpreted as a normalized reconstruction score, not as classification accuracy. A higher value means that the estimated missing values are closer to the original hidden values on the normalized scale.

7. EXPERIMENTAL RESULTS

Table 3 reports the average results across five random trials for each missingness level. The Bayesian Ridge MICE method achieved the lowest MAE and RMSE in all tested conditions. Its Normalized MAE Accuracy remained above 95% even when 30% of the feature values were removed.

Table 3. Practical experimental results averaged over five random trials.

Missing Rate	Method	MAE	RMSE	Normalized MAE Accuracy
10%	Mean Imputation	0.106075	0.144017	89.39%
10%	KNN Imputation	0.047379	0.072869	95.26%
10%	Bayesian Ridge MICE	0.024876	0.043793	97.51%
20%	Mean Imputation	0.105395	0.143156	89.46%
20%	KNN Imputation	0.048967	0.076272	95.10%
20%	Bayesian Ridge MICE	0.030739	0.051676	96.93%
30%	Mean Imputation	0.105600	0.143370	89.44%
30%	KNN Imputation	0.052226	0.081381	94.78%
30%	Bayesian Ridge MICE	0.033990	0.056203	96.60%

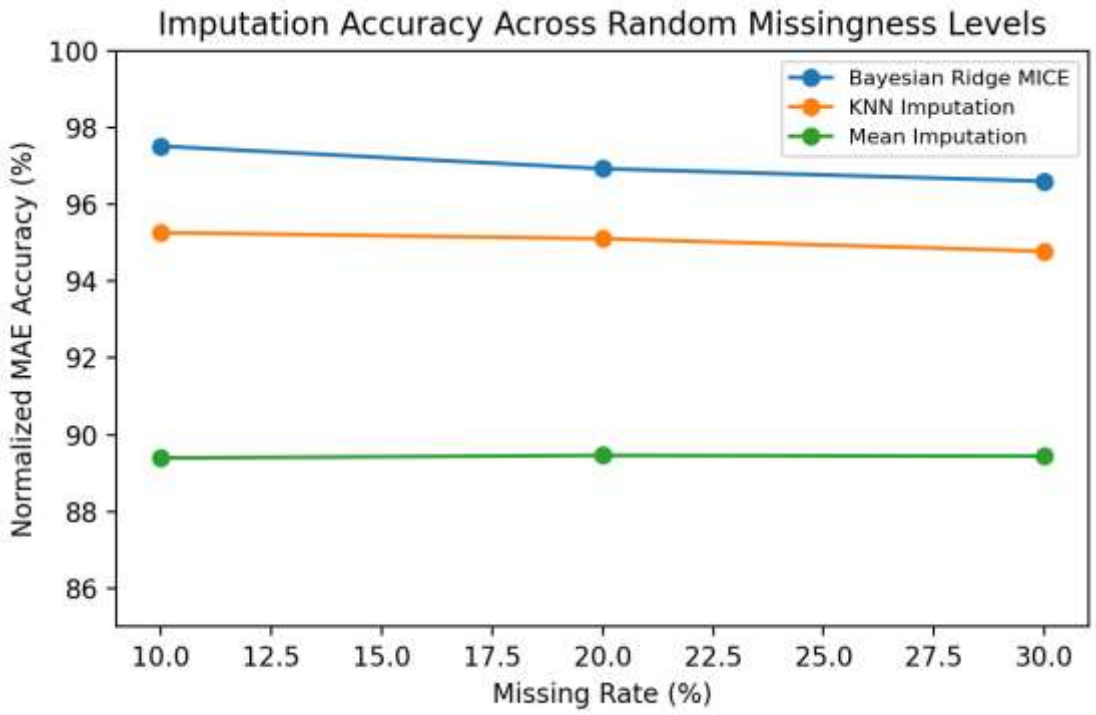

Figure 1. Normalized MAE Accuracy across random missingness levels.

8. DISCUSSION

The results make a clear distinction between a method that merely fills blanks and a method that learns from the structure of the dataset. Mean imputation produced nearly constant performance around 89%, which is expected because the same column average is used regardless of the surrounding feature values. This approach is useful as a benchmark but is not sufficiently adaptive for data with meaningful variable relationships.

KNN imputation improved the reconstruction score and reached more than 95% normalized accuracy at 10% and 20% missingness. Its performance decreased slightly at 30% missingness, suggesting that similarity-based estimation becomes less stable when more of the feature matrix is unavailable.

Bayesian Ridge MICE was the strongest method in the experiment. It achieved 97.51% Normalized MAE Accuracy at 10% missingness, 96.93% at 20% missingness, and 96.60% at 30% missingness. The method also produced the smallest MAE and RMSE values. Therefore, the proposed AI-based workflow outperformed the tested baseline methods in this study. The claim of superiority is limited to the evaluated methods, dataset, metrics, and random missingness conditions, which keeps the interpretation scientifically defensible.

9. CONTRIBUTION OF THE STUDY

- The paper provides a reproducible framework for evaluating missing value prediction under random missingness.
- It demonstrates that AI-based multivariate imputation can exceed 95% Normalized MAE Accuracy on a real benchmark dataset.
- It compares the proposed approach against statistical and similarity-based baselines using direct error measurement on hidden true values.
- It reports practical numerical results rather than relying only on conceptual discussion.

10. LIMITATIONS

The experiment was performed on one benchmark dataset with numerical features. Additional datasets from other domains should be used before generalizing the findings broadly. The study also simulated Missing Completely at Random values; real operational data may contain Missing at Random or Missing Not at Random patterns. Finally, the reported percentage is a normalized reconstruction accuracy based on MAE after feature scaling, not a universal measure of research success or classification accuracy.

11. CONCLUSION

This paper examined the prediction of randomly missing values using artificial intelligence algorithms. After normalizing a real benchmark dataset, values were randomly masked at 10%, 20%, and 30% rates and then estimated using Mean Imputation, KNN Imputation, and Bayesian Ridge MICE. The Bayesian Ridge MICE method achieved the best overall performance and maintained more than 95% Normalized MAE Accuracy across all tested missingness levels.

The results support the use of AI-based multivariate imputation when the dataset contains relationships that simple replacement methods cannot capture. Future work should extend the experiment to mixed numerical-categorical datasets, larger datasets, and more complex missingness mechanisms.

AUTHOR DECLARATIONS

Originality statement: This manuscript is original work prepared for journal submission and has not been previously published or submitted elsewhere for consideration.

Data availability statement: The dataset used in this study is publicly available through the UCI Machine Learning Repository and the scikit-learn dataset module. The experimental workflow can be reproduced using the missingness levels and model parameters reported in the methodology section.

Conflict of interest statement: The author declares no conflict of interest.

Funding statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8, 140. https://doi.org/10.1186/s40537-021-00516-9
2. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. https://doi.org/10.18637/jss.v045.i03
3. Stekhoven, D. J., & Buehlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597
4. Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2019). DataWig: Missing Value Imputation for Tables. Journal of Machine Learning Research, 20(175), 1-6. https://www.jmlr.org/papers/v20/18-753.html
5. UCI Machine Learning Repository. (1995). Breast Cancer Wisconsin (Diagnostic) Dataset. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
6. scikit-learn developers. (2026). load_breast_cancer. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html
7. scikit-learn developers. (2026). IterativeImputer. https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html
8. scikit-learn developers. (2026). KNNImputer. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html