International Journal of Mathematics and Statistics Studies, 14 (1),68-79, 2026
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Enhancing Predictive Accuracy in Incomplete Datasets: A Comparative Evaluation of Missing-Data Imputation Methods

Jude C. Obi and Ebere J. Odubuasi
Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Uli, Nigeria
Correspondence: Jude C. Obi (oj.obi@coou.edu.ng); Ebere J. Odubuasi (ebereodubuasi@yahoo.com)

doi: https://doi.org/10.37745/ijmss.13/vol14n16879 Published September 06, 2026

Citation: Jude C. Obi and Ebere J. Odubuasi (2026) Enhancing Predictive Accuracy in Incomplete Datasets: A Comparative Evaluation of Missing-Data Imputation Methods, *International Journal of Mathematics and Statistics Studies*, 14 (1),68-79

Abstract: *Missing data remains one of the most persistent challenges in statistical modelling and machine learning, often degrading predictive accuracy and leading to biased inference when handled inappropriately. Although numerous imputation techniques have been developed, their comparative effectiveness under different missingness mechanisms remains an active area of research. This study comparatively evaluates four widely used imputation techniques: Mean Imputation, k-nearest neighbours (KNN), Multiple Imputation by Chained Equations (MICE), and Autoencoder-based Imputation, with respect to their ability to preserve predictive performance across multiple benchmark datasets. Seventeen datasets representing diverse application domains were employed. Artificial missingness was introduced under three missing-data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Following imputation, Random Forest models were fitted to the reconstructed datasets, and predictive performance was evaluated using the Root Mean Squared Error (RMSE). The original complete datasets served as the benchmark against which the performance of each imputation method was assessed. Comparative evaluation was conducted using absolute deviation from the benchmark RMSE, paired hypothesis tests, Bonferroni-adjusted multiple comparisons, and non-parametric tests where appropriate. The results demonstrate that Multiple Imputation consistently produced RMSE values closest to those obtained from the original complete datasets under all three missingness mechanisms. KNN imputation generally ranked second, whereas Mean Imputation and Autoencoder-based Imputation exhibited larger deviations from the benchmark, particularly under more challenging missingness conditions. Statistical analyses further showed that Multiple Imputation most consistently preserved the predictive characteristics of the original datasets, while the remaining methods displayed varying levels of performance depending on the missingness mechanism and dataset characteristics. The findings highlight the importance of selecting imputation methods according to both the missing-data mechanism and the intended predictive modelling objective. Overall, Multiple Imputation provides the most reliable and robust performance across heterogeneous datasets, making it an appropriate general-purpose imputation strategy for predictive analytics involving incomplete data.*

Keywords: missing data, multiple imputation, K-nearest neighbours, autoencoder, predictive modelling, random forest, root mean squared error

International Journal of Mathematics and Statistics Studies, 14 (1),68-79, 2026
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

INTRODUCTION

The rapid growth of data-driven decision-making has transformed scientific research, industrial processes, healthcare delivery, financial forecasting, engineering design, and public policy. Across these diverse application areas, predictive models increasingly support critical decisions whose reliability depends fundamentally on the quality and completeness of the underlying data. Unfortunately, missing observations are almost unavoidable in real-world datasets. Data may become incomplete because of non-response during surveys, equipment malfunction, transmission failures, human recording errors, sensor outages, administrative omissions, or privacy-related restrictions. Consequently, missing data remains one of the most important challenges confronting statisticians, data scientists, and machine learning practitioners.

Incomplete data can substantially influence both statistical inference and predictive modelling. Ignoring missing observations or handling them inappropriately may reduce statistical efficiency, distort relationships among variables, bias parameter estimates, and ultimately degrade predictive accuracy. These effects become increasingly pronounced when the proportion of missing observations increases or when the missingness mechanism depends systematically on observed or unobserved characteristics of the data. Consequently, the selection of an appropriate imputation strategy has become a fundamental component of modern statistical analysis.

Early approaches to handling missing data relied primarily on complete-case analysis, listwise deletion, pairwise deletion, and simple replacement methods such as mean substitution. Although computationally inexpensive and straightforward to implement, these approaches frequently underestimated data variability, weakened correlations among variables, and ignored the uncertainty associated with missing observations. As a result, they may produce misleading statistical conclusions and inferior predictive performance, particularly when the proportion of missing data is moderate or substantial.

Advances in computational statistics and machine learning have led to the development of increasingly sophisticated imputation techniques capable of exploiting complex relationships within datasets. Among the most widely adopted methods are k-nearest neighbours (KNN) imputation, which estimates missing values using neighbouring observations; Multiple Imputation by Chained Equations (MICE), which generates multiple plausible values while accounting for uncertainty; and deep learning approaches such as Autoencoder-based Imputation, which reconstruct incomplete observations through nonlinear representation learning. These techniques have demonstrated improved performance over conventional methods in many application domains, although their effectiveness often depends on the structure of the data and the underlying missingness mechanism.

The theoretical foundation for missing-data analysis is largely based on Rubin's classification of missingness mechanisms into Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Under MCAR, the probability of missingness is independent of

both observed and unobserved values. MAR assumes that missingness depends only on observed variables, whereas MNAR arises when the probability of missingness depends directly on the missing values themselves or on other unobserved factors. Since these mechanisms affect the amount of information available for imputation, no single algorithm is expected to perform optimally under all circumstances. Comparative evaluation across multiple missingness mechanisms therefore provides a more realistic assessment of algorithm performance than evaluation under a single scenario.

Despite the growing literature on missing-data imputation, several limitations remain. Many previous investigations have focused on individual algorithms or have compared only a limited number of methods using a small collection of datasets. Other studies have concentrated primarily on imputation accuracy without adequately examining the influence of imputation on downstream predictive modelling. In practice, however, the ultimate objective of data imputation is not merely to estimate missing values accurately but to preserve the predictive characteristics of the original dataset. Consequently, evaluating imputation methods using predictive performance provides a more meaningful assessment of their practical usefulness.

This study addresses these limitations through a comprehensive benchmark-based comparison of four widely used imputation techniques: Mean Imputation, KNN imputation, Multiple Imputation by Chained Equations, and Autoencoder-based Imputation. Seventeen benchmark datasets drawn from multiple application domains are considered to ensure that the findings are not restricted to a single problem type. Artificial missingness is introduced under MCAR, MAR, and MNAR mechanisms, after which each imputation technique is applied using a standardized analytical framework. Predictive performance is subsequently evaluated using Random Forest models and Root Mean Squared Error (RMSE), with the original complete datasets serving as the performance benchmark.

Unlike many previous comparative studies, the present research evaluates imputation techniques according to how closely they preserve the predictive performance of the original complete datasets rather than solely on the accuracy of the imputed values themselves. This benchmark-oriented perspective provides practical guidance for researchers and practitioners seeking imputation methods that maintain the predictive integrity of incomplete datasets.

The specific objectives of this study are to:

(a) Compare the predictive performance of Mean Imputation, KNN imputation, Multiple Imputation, and Autoencoder-based Imputation across seventeen benchmark datasets;

(b) Evaluate the influence of MCAR, MAR, and MNAR missingness mechanisms on the effectiveness of each imputation method;

(c) Determine which imputation technique most closely reproduces the predictive performance of the original complete datasets; and

(d) Provide evidence-based recommendations for selecting appropriate imputation methods for predictive modelling involving incomplete data.

The remainder of this paper is organised as follows. Section 2 describes the methodology, including the experimental design, benchmark datasets, missingness generation procedures, imputation algorithms, and statistical evaluation methods. Section 3 presents the empirical results together with their interpretation and comparative discussion. Section 4 concludes the paper by summarizing the principal findings, highlighting the limitations of the study, and suggesting directions for future research.

METHODOLOGY

Research Design

This study adopted a comparative experimental research design to evaluate the effectiveness of four missing-data imputation techniques in preserving predictive accuracy. The experimental approach was selected because it permits the systematic comparison of competing imputation methods under controlled missing-data conditions while maintaining identical predictive modelling procedures for all datasets.

The experimental framework comprised five sequential stages. First, complete benchmark datasets were obtained from publicly available repositories. Second, artificial missing values were introduced according to three established missing-data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Third, each incomplete dataset was imputed independently using four alternative imputation techniques. Fourth, Random Forest regression models were trained using each imputed dataset. Finally, predictive performance was evaluated by comparing the Root Mean Squared Error (RMSE) obtained from each imputed dataset with the RMSE obtained from the corresponding original complete dataset, which served as the benchmark.

Using the original dataset as the performance benchmark provides a practical basis for determining how effectively each imputation method preserves the predictive characteristics of completed data rather than merely reconstructing missing values.

Benchmark Datasets

Seventeen publicly available benchmark datasets were employed in this study. The datasets were selected because they represent diverse application domains and exhibit substantial variation in

dimensionality, sample size, and predictor characteristics. This diversity enable comprehensive evaluation of imputation methods under heterogeneous data conditions and improves the generalisability of the findings.

The benchmark datasets included Red Wine Quality, White Wine Quality, air quality, wireless sensor network (WSN), Concrete Compressive Strength, protein structure, superconductivity, synchronous machine, together with nine additional benchmark regression datasets obtained from established machine learning repositories. These datasets collectively represent applications in engineering, environmental science, manufacturing, agriculture, transportation, and industrial systems.

Prior to introducing missing values, all datasets were screened for completeness, duplicate observations, inconsistent variable types, and obvious recording errors. Continuous predictor variables were maintained in their original measurements scales throughout the benchmarking procedure.

Artificial Generation of Missing Data

To ensure objective comparison across imputation techniques, artificial missing values were generated from the complete datasets rather than using naturally incomplete data. Three missing-data mechanisms were considered:

Missing Completely at Random (MCAR): Missing observations were introduced independently of both observed and unobserved variables, thereby satisfying the MCAR assumption.

Missing at Random (MAR): Missingness was generated conditional on observed variables using logistic probability models, allowing the probability of deletion to depend on available information.

Missing Not at Random (MNAR): Missing observations were generated conditional on the variable containing the missing values, representing situations where missingness depends directly on the unobserved values.

The missing-data mechanisms were implemented in the R statistical environment using the missMethods package, which provides reproducible procedures for generating MCAR, MAR and MNAR data structures through probabilistic deletion mechanisms. Random seeds were fixed before each simulation to ensure complete reproducibility of all experimental results.

Imputation Methods

Four imputation techniques representing increasing methodological complexity were evaluated.

Mean Imputation

Mean Imputation replaces every missing observation by the arithmetic mean of the corresponding variable,

$$\hat{x}_{ij} = \bar{x}_j$$

Where (\bar{x}_{j}) denotes the mean of the observed values in variable (j). Although computationally simple, this method ignores relationships among variables and typically underestimates data variability.

K-Nearest Neighbours (KNN) Imputation

KNN imputation estimates missing observations using the values of the (k) most similar observations determined through a distance metric. Missing values are replaced by the average of neighbouring observations,

$$\hat{x}_{ij} = (1/k) \Sigma_{r=1}^{k} x_{rj}$$

The method exploits local data structure and often performs well when similar observations exist within the dataset.

Multiple Imputation by Chained Equations (MICE)

Multiple Imputation generates several plausible values for each missing observation using iterative conditional regression models. Each completed dataset is analysed independently before parameter estimates are combined according to Rubin's combining rules. Unlike single imputation techniques, MICE incorporates uncertainty associated with missing values, thereby producing statistically valid inference.

Autoencoder-Based Imputation

Autoencoder Imputation employs an artificial neural network trained to reconstruct incomplete observations through nonlinear feature learning. The encoder compresses the input data into a lower-dimensional latent representation, while the decoder reconstructs the original observations. Missing values are estimated from the reconstructed output, enabling the model to capture complex nonlinear relationships among predictor variables.

Predictive Modelling

Following imputation, each completed dataset was analysed using a Random Forest regression model. Random Forest was selected because of its robustness, ability to model nonlinear relationships, resistance to overfitting, and strong predictive performance across diverse benchmark datasets.

For each experiment, the same modelling procedure was maintained across all imputation methods to ensure that observed differences in predictive performance could be attributed solely to the imputation strategy rather than variations in model specification.

The predictive performance obtained from the original complete dataset served as the benchmark against which all imputed datasets were evaluated.

Performance Evaluation

Predictive accuracy was assessed using the Root Mean Squared Error (RMSE),

$$RMSE = \sqrt{[(1/n) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2]}$$

Where (y_i) denotes the observed response and (\hat{y}_i) represents the corresponding model prediction.

To evaluate how closely each imputation method reproduced the predictive performance of the complete dataset, the absolute deviation from the benchmark RMSE was calculated as

$$AE_{ij} = |RMSE_{ij} - RMSE_i^{Original}|$$

Where (AE_{ij}) denotes the benchmark deviation for dataset (i) under imputation method (j).

Smaller absolute deviations indicate greater similarity to the predictive performance of the original complete dataset.

Statistical Analysis

Descriptive and inferential statistical procedures were employed to compare the four imputation techniques.

First, benchmark deviation measures, including the mean absolute deviation, median absolute deviation and RMSE of deviations, were computed to quantify the proximity of each imputation method to the original datasets.

Second, paired hypothesis tests were performed by comparing the RMSE obtained from each imputation method with the benchmark RMSE derived from the corresponding original dataset. Where appropriate, paired t-tests were complemented by Wilcoxon signed-rank tests to provide robustness against departures from normality.

To control the family-wise error rate arising from multiple pairwise comparisons, Bonferroni-adjusted significance levels were applied. In addition, an overall Friedman test was used to examine whether statistically significant differences existed among the competing imputation methods across the benchmark datasets. The analyses were conducted at the 5% level of significance.

Software Environment

All analyses were performed using the R statistical computing environment. Missing-data generation, imputation, predictive modelling and statistical analyses were implemented through standard R packages, while random seeds were specified to ensure reproducibility of the experimental results.

RESULTS AND DISCUSSION

Performance under Missing Completely at Random (MCAR)

The performance of the four imputation techniques under the MCAR mechanism was evaluated by comparing the RMSE obtained from each imputed dataset with that of the corresponding original complete dataset. The original dataset served as the benchmark, while the absolute deviation from the benchmark RMSE was used to quantify the extent to which each imputation method preserved predictive performance.

The results indicate that Multiple Imputation consistently produced the smallest deviation from the benchmark RMSE across the seventeen datasets. It achieved the lowest mean absolute deviation (1.1559) and median absolute deviation (0.3335), demonstrating the closest agreement with the

predictive performance of the original datasets. KNN imputation ranked second, whereas autoencoder and Mean Imputation exhibited substantially larger deviations from the benchmark.

A dataset-level comparison further strengthened these findings. Multiple Imputation produced the smallest deviation from the benchmark in eight of the seventeen datasets, while KNN achieved this distinction in six datasets. Autoencoder and Mean Imputation were the closest methods in only a few cases, indicating less consistent performance across heterogeneous datasets.

Paired hypothesis tests revealed that the RMSE produced by Multiple Imputation was not significantly different from the benchmark dataset, whereas Mean Imputation and Autoencoder Imputation exhibited statistically significant departures from the original data. KNN imputation produced mixed evidence, suggesting that although its overall performance was competitive, its effectiveness depended on dataset characteristics.

These findings suggest that Multiple Imputation effectively preserves the statistical structure of complete datasets under purely random missingness. By incorporating uncertainty into the estimation process, MICE avoids the systematic bias commonly associated with deterministic single-imputation methods. In contrast, Mean Imputation ignores multivariate relationships and consequently introduces greater distortion into predictive modelling. Although Autoencoder Imputation is capable of modelling nonlinear relationships, its performance under MCAR appears less stable, possibly because randomly distributed missing values provide limited structural information for effective neural network reconstruction.

Performance under Missing at Random (MAR)

Under the MAR mechanism, the comparative performance of the four imputation methods remained largely consistent with the MCAR results. Multiple Imputation again demonstrated the smallest overall deviation from the benchmark RMSE, with KNN imputation closely following. Mean Imputation and Autoencoder Imputation produced substantially larger deviations, indicating reduced ability to preserve the predictive characteristics of the complete datasets.

The paired hypothesis tests indicated that none of the imputation methods differed significantly from the benchmark after Bonferroni adjustment, despite some methods appearing statistically different before adjustment. These findings demonstrate the importance of controlling the family-wise error rate when multiple pairwise comparisons are performed.

Overall, Multiple Imputation achieved the best average performance, while KNN exhibited the smallest median deviation from the benchmark, suggesting that both methods effectively preserved predictive accuracy under MAR conditions. Nevertheless, Multiple Imputation demonstrated greater consistency across the full collection of benchmark datasets, supporting its stronger overall benchmark performance in this experiment.

The superior performance of Multiple Imputation under MAR is theoretically expected because the method explicitly models the relationships between variables when estimating missing values. Since the MAR mechanism assumes that missingness depends only on observed information, MICE is well

suited to exploit this dependency. KNN also performed favourably because neighbouring observations often contain sufficient information to recover missing values under MAR. In contrast, Mean Imputation ignores these relationships, while autoencoder performance appears to be influenced by the heterogeneity of the benchmark datasets.

Performance under Missing Not at Random (MNAR)

The MNAR mechanism presented the greatest challenge for all four imputation methods because the probability of missingness depended directly on the missing values themselves. Consequently, every imputation method exhibited larger deviations from the benchmark than under MCAR and MAR.

Despite the increased difficulty, Multiple Imputation again achieved the smallest mean absolute deviation from the benchmark RMSE, followed closely by KNN imputation. Mean Imputation produced the largest overall deviation, whereas Autoencoder Imputation demonstrated intermediate performance. Multiple Imputation also achieved the highest number of dataset-level wins, indicating comparatively robust benchmark performance under the most challenging missingness scenario.

The paired hypothesis tests indicated no statistically significant difference between the benchmark RMSE and the four imputation methods when assessed using the paired t-test. However, the Wilcoxon signed-rank test identified a significant difference for Multiple Imputation, reflecting the influence of non-normality and several extreme datasets.

These findings illustrate the increased complexity associated with imputing MNAR data. Since missingness depends on unobserved values, no imputation method can completely recover the original data-generating process. Nevertheless, Multiple Imputation continued to outperform the competing methods in terms of overall proximity to the benchmark, suggesting that modelling uncertainty remains advantageous even under challenging missingness conditions.

Comparative Discussion

A comparison of the three missingness mechanisms reveals a remarkably consistent pattern. Across MCAR, MAR, and MNAR, Multiple Imputation repeatedly demonstrated the closest agreement with the predictive performance of the original complete datasets. KNN imputation consistently ranked second, while Mean Imputation generally produced the least satisfactory results. Autoencoder Imputation yielded mixed performance, performing competitively for selected datasets but lacking the consistency required to outperform Multiple Imputation across heterogeneous benchmark datasets. These findings demonstrate that preserving predictive accuracy depends not only on recovering missing observations but also on maintaining the multivariate relationships among variables. Multiple Imputation appears to achieve this objective more effectively than the competing techniques because it explicitly accounts for uncertainty and variable interdependence during the imputation process. KNN provides a computationally attractive alternative with competitive performance, particularly when local neighbourhood structures are well defined. Conversely, Mean Imputation sacrifices statistical fidelity for computational simplicity, while Autoencoder-based Imputation may require substantially larger datasets and careful hyperparameter optimisation before its theoretical advantages can be fully realised.

From a practical perspective, the findings suggest that researchers undertaking predictive modelling with incomplete datasets may consider Multiple Imputation as a strong default candidate when computational resources permit. KNN imputation represents a suitable alternative for moderately sized datasets, whereas Mean Imputation should be used with caution because of its tendency to distort predictive performance. Although autoencoder-based methods continue to attract considerable attention in machine learning, the present benchmark results indicate that they do not consistently outperform established statistical approaches across diverse real-world datasets.

Conclusion and Recommendations

This study conducted a comprehensive comparative evaluation of four missing-data imputation techniques—Mean Imputation, K-nearest neighbours (KNN), Multiple Imputation by Chained Equations (MICE), and Auto encoder-based Imputation, using seventeen benchmark datasets representing diverse application domains. Artificial missing values were generated under the three classical missing-data mechanisms, namely Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The predictive performance of each imputation method was assessed using Random Forest models, with the Root Mean Squared Error (RMSE) of the original complete datasets serving as the benchmark.

The empirical results consistently demonstrated that Multiple Imputation most effectively preserved the predictive characteristics of the original datasets across the three missingness mechanisms. In particular, Multiple Imputation produced the smallest average deviation from the benchmark RMSE and exhibited the greatest consistency across the seventeen benchmark datasets. KNN imputation generally ranked second and provided competitive performance, especially under MCAR and MAR conditions. In contrast, Mean Imputation consistently produced the largest departures from the benchmark, confirming its limited suitability for predictive modelling involving incomplete data. Although Autoencoder-based Imputation showed competitive performance for some individual datasets, its overall performance was less consistent than that of Multiple Imputation.

The inferential analyses further supported these findings. Paired hypothesis tests and Bonferroni-adjusted multiple comparisons indicated that Multiple Imputation most closely reproduced the predictive behaviour of the original complete datasets. The benchmark-base devaluation adopted in this study therefore suggests that preserving predictive accuracy benefits from imputation methods capable of modelling the multivariate relationships and uncertainty inherent in incomplete data rather than relying on simple deterministic replacement strategies.

The findings of this study have important practical implications for researchers and practitioners working with incomplete datasets. For general predictive modelling applications, Multiple Imputation emerged as the strongest general-purpose candidate among the methods evaluated because of its robustness, statistical validity, and consistent predictive performance across diverse datasets. KNN imputation provides a suitable alternative where computational simplicity or implementation constraints are important considerations. Conversely, Mean Imputation should be used cautiously because of its tendency to distort predictive performance. While autoencoder-based methods remain

promising, their application may require larger datasets, careful hyperparameter optimisation, and greater computational resources before their theoretical advantages can be fully realised.

Although this study considered multiple benchmark datasets and three missing-data mechanisms, certain limitations should be acknowledged. The comparative evaluation was restricted to four imputation technique stand-alone single predictive learning algorithm. Other modern imputation approaches, including ensemble learning and transformer-based methods, were beyond the scope of the present investigation. Furthermore, the study focused exclusively on predictive accuracy measured by RMSE; additional evaluation metrics may provide complementary insights into imputation performance. In addition, the present manuscript reports the common analytical framework but does not provide all implementation-level tuning details for every imputation and Random Forest fit. Accordingly, the findings should be interpreted as benchmark evidence within the stated experimental design, and future replications should report complete hyperparameter settings, data-splitting procedures, missingness proportions, and software/package versions.

Future studies may extend the present work by incorporating larger collections of benchmark datasets, alternative predictive learning algorithms, additional missing-data scenarios, and recently developed deep-learning imputation methods. Comparative investigations involving classification problems, high-dimensional data, longitudinal datasets, and real-world clinical or financial applications would further enhance understanding of the practical effectiveness of modern missing-data imputation techniques.

In conclusion, this study demonstrates that the choice of imputation method substantially influences predictive performance. Among the methods investigated, Multiple Imputation provides the most reliable and consistent approximation to the predictive performance of complete datasets, making it the preferred general-purpose approach for predictive modelling in the presence of missing data.

REFERENCES

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.

Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data (3rd ed.). Wiley.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525.

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In Proceedings of the 35th International Conference on Machine Learning.

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5–6), 519–533.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115.

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), 263–282.

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1), 1–31.

van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). CRC Press.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). The Elements of Statistical Learning (2nd ed., corrected printing). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.