A Review of Missing Data Handling Techniques in Data Science

Aishwarya Kolhe[1], Geeta Yadav[2], Dnyaneshwari Chalak[3], Dr. Shital Ghotekar[4]
[1,2,3,4]*Department of Science & Computer Science, MAEER's MIT Arts, Commerce and Science College, Alandi (D), Pune, India*
doi.org/10.64643/ijirt.208754-459

***Abstract*—Incomplete data is one of the most common quality problems encountered in data science. Missing values can occur because of non-response, faulty sensors, data-entry mistakes, transmission failures, privacy restrictions, or differences between data sources. If missing values are handled without considering the reason for missingness, the resulting analysis may be biased and machine-learning models may lose predictive reliability. This paper presents a review of major approaches for handling missing data, beginning with three classical missingness mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [1], [2]. The reviewed methods include deletion, statistical imputation, regression-based imputation, multiple imputation, expectation-maximization, K-nearest-neighbour imputation, random-forest-based methods, and deep-learning approaches [4][10]. These techniques are compared in terms of their assumptions, information retention, computational requirements, scalability, and suitability for different data conditions. The review emphasizes that the effectiveness of an imputation method depends strongly on the mechanism and pattern of missingness. Simple methods may be suitable for small amounts of approximately random missingness, whereas advanced approaches can be useful when variables have strong relationships or the dataset is high-dimensional. Important challenges include uncertainty about the missingness mechanism, computational cost, evaluation when the actual values are unavailable, and the possibility of introducing or amplifying bias. The paper concludes by identifying future directions involving adaptive, uncertainty-aware, scalable, and fairness-conscious missing-data frameworks.**

***Index Terms*—Missing Data, Data Imputation, MCAR, MAR, MNAR, Data Preprocessing, Machine Learning.**

I. INTRODUCTION

Data driven systems depend heavily on the quality of the information supplied to them. In real world datasets, however, complete observations are uncommon. A healthcare record may lack a laboratory measurement, a survey may contain unanswered questions, a sensor may stop transmitting for a particular period, and a business database may contain fields that were not collected for every customer. These situations result in missing values that must be considered before statistical analysis or machine learning modelling.

Missing data is not simply a formatting problem. The way missing observations are removed or replaced can change distributions, correlations, and relationships among variables. For predictive models, inappropriate preprocessing can result in information loss and systematic errors. Therefore, missing data handling should be considered a methodological decision rather than a routine preprocessing operation [2], [10].

A central concept in missing-data analysis is that the reason for missingness matters. Rubin introduced a framework that distinguishes Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [1]. These mechanisms provide an important basis for selecting an appropriate missing data treatment.

For example, deletion may be appropriate in limited situations where missingness is approximately MCAR, while model-based and multiple-imputation techniques are often considered under MAR assumptions [2], [4]. MNAR situations are more difficult because the probability of missingness may depend on the unobserved value itself. This paper reviews widely used approaches, ranging from deletion and simple statistical imputation to machine learning and deep learning techniques. The objective is not to identify one universally superior method but to compare the major approaches according to their assumptions, advantages, limitations, computational requirements, and possible applications.

II. MISSING DATA MECHANISMS

The mechanism responsible for missing values influences both the potential bias of an analysis and the suitability of an imputation method. The three classical categories are MCAR, MAR, and MNAR [1], [2].

A. Missing Completely at Random (MCAR)

Under MCAR, the probability that a value is missing does not depend on either observed or unobserved values. In an ideal MCAR situation, records containing missing observations are not systematically different from complete records with respect to the variables being analysed [1].

When the amount of missingness is small, deletion may therefore have a relatively limited effect. However, strict MCAR conditions are difficult to establish in many practical datasets. Consequently, researchers should avoid automatically assuming MCAR simply because the missing-data percentage is small.

B. Missing at Random (MAR)

Under MAR, missingness can depend on information that is already observed, but after accounting for those observed variables, the probability of missingness does not depend on the missing value itself [2].

For example, missing income information may be related to an individual's observed age, occupation, or employment category. Multiple-imputation procedures such as MICE can be applied under assumptions related to MAR [4].

MAR is an important assumption in many statistical approaches to missing-data analysis. However, determining whether MAR is appropriate cannot generally be confirmed from the observed data alone.

C. Missing Not at Random (MNAR)

MNAR occurs when the probability that a value is missing is related to the unobserved value itself, even after accounting for available information [1], [2].

For example, individuals may be less likely to report a sensitive attribute because of the actual value of that attribute. MNAR is particularly challenging because the missingness process contains information that is not directly available in the observed dataset.

In such situations, simple imputation may be inappropriate. Sensitivity analysis or explicit modelling of the missingness process may be required.

III. RELATED WORK

Missing-data methodology has developed from classical statistical methods to flexible machine-learning approaches. Rubin established a formal framework for understanding missing-data mechanisms and inference [1]. Little and Rubin provided a comprehensive treatment of statistical analysis involving incomplete data, including deletion, likelihood-based methods, and imputation techniques [2].

Schafer studied multivariate approaches for incomplete datasets and highlighted the importance of modelling relationships among variables when estimating missing observations [3].

Multiple imputation became an important approach because it incorporates uncertainty rather than treating a single estimated value as completely known. Van Buuren and Groothuis-Oudshoorn introduced the widely used MICE framework for multivariate imputation by chained equations [4].

Machine-learning-based approaches have also been investigated. MissForest uses random forests to estimate missing values iteratively and can capture nonlinear relationships between variables [5]. K nearest neighbour approaches estimate missing values using observations with similar characteristics [6].

More recent research has examined neural network and generative approaches. GAIN uses generative adversarial networks for missing data imputation and attempts to learn complex relationships within incomplete datasets [7]. The Expectation Maximization algorithm provides another important likelihood-based approach for incomplete data [8].

Research comparing missing data treatment methods has shown that the performance of an approach depends on the missingness mechanism, missingness percentage, data characteristics, and downstream learning task [9], [10].

IV. MAJOR MISSING DATA HANDLING TECHNIQUES

A. Deletion-Based Methods

Listwise deletion removes an entire observation when one or more required fields are missing. It is simple, fast, and easy to implement. However, it can considerably reduce the available sample size.

Deletion may also introduce bias when missingness is associated with characteristics of the observations. Therefore, it is generally more suitable when the amount of missing data is small and the missingness can reasonably be considered random [2].

Pairwise deletion uses all available pairs of observations for particular statistical calculations. Although this can preserve more data than listwise deletion, different calculations may be based on different subsets of observations, which can lead to inconsistencies.

B. Single Imputation

Mean, median, and mode imputation replace missing values with simple summary statistics. These methods require little computation and are easy to understand.

Mean imputation can be used for numerical variables, while median imputation may be preferable for skewed distributions. Mode imputation is commonly used for categorical variables.

The major disadvantage is that these methods reduce natural variability by inserting repeated values. They may also distort correlations and underestimate uncertainty.

Regression imputation predicts missing values using other observed variables. Because it considers relationships among variables, it can preserve more structure than simple mean or median replacement. However, deterministic regression imputation may underestimate uncertainty because the predicted value is treated as if it were known with certainty.

C. Multiple Imputation

Multiple imputation generates several plausible completed datasets instead of replacing a missing value with a single estimate [2], [4].

Each completed dataset is analysed separately, and the resulting estimates are combined. This approach accounts for both the variation within individual imputations and the variation between imputations.

The MICE framework is a widely used implementation of multiple imputation by chained equations [4].

Multiple imputation can provide a more statistically appropriate treatment of uncertainty than single imputation, although it requires additional computation and implementation effort.

D. Expectation-Maximization

The Expectation-Maximization (EM) algorithm is an iterative likelihood-based approach for incomplete data [8].

The algorithm consists mainly of two stages. In the Expectation step, the missing observations are estimated using current parameter values. In the Maximization step, the model parameters are updated based on the estimated complete data information.

These steps are repeated until the algorithm reaches convergence. EM is useful when the underlying probabilistic assumptions are appropriate, but its performance depends on the selected statistical model and can become computationally expensive.

E. K-Nearest Neighbour Imputation

K-nearest neighbour (KNN) imputation estimates a missing value by identifying observations that are most similar to the incomplete observation [6].

For numerical variables, similarity can be determined using distance measures. The missing value can then be estimated using the values of the selected nearest neighbours.

KNN can capture local relationships that simple statistical methods cannot. However, its performance depends on feature scaling, distance measurement, and the selected value of k. It can also become computationally expensive for large and high dimensional datasets.

F. Machine Learning Based Imputation

Machine learning techniques treat missing value estimation as a prediction problem. Random forest-based methods such as MissForest iteratively predict missing values using other variables in the dataset [5].

Random forests can model nonlinear relationships and can handle mixed numerical and categorical data.

This flexibility makes them useful for complex datasets.

However, machine learning methods generally require more computational resources than simple imputation. Model complexity can also make the results more difficult to interpret and validate.

G. Deep-Learning Approaches

Deep learning methods can learn complex representations from incomplete datasets. Neural network-based methods such as autoencoders can be trained to reconstruct missing components.

GAIN uses generative adversarial networks for missing-data imputation [7]. Generative approaches can capture complex dependencies between variables and may be useful for high-dimensional data.

However, deep learning methods usually require larger datasets, greater computational resources, and careful hyperparameter selection. Their complexity can also make the imputation process less transparent.

V. COMPARATIVE ANALYSIS

The major missing-data handling approaches can be compared according to assumptions, information retention, computational cost, and suitable applications.

Technique	Typical Assumption	Information Retention	Cost	Suitable Situation
Listwise deletion	Mainly MCAR	Low	Very Low	Small, approximately random missingness
Mean/Median/Mode	Simple MCAR/MAR use	Low–Moderate	Very Low	Quick baseline preprocessing
Regression	MAR-related	Moderate	Low–Moderate	Strong feature relationships
Multiple Imputation	Usually, MAR	High	Moderate–High	Statistical analysis with uncertainty
EM	Model-dependent	High	Moderate–High	Suitable probabilistic models
KNN	Relationship-based	Moderate–High	High at scale	Strong local similarity
Random Forest	Flexible/model-based	High	Moderate–High	Nonlinear, mixed-type data
Deep Learning	Data/model dependent	High	Very High	Large, complex datasets

The comparison indicates that no single method is universally superior. Deletion has very low computational cost but can discard useful observations. Mean and median imputation are simple but may reduce natural variability.

Multiple imputation is useful when uncertainty is important [2], [4]. KNN and random forest approaches can capture local and nonlinear relationships [5], [6]. Deep learning methods provide additional flexibility for complex datasets but require greater computational resources [7], [10]. Therefore, method selection should consider the missingness mechanism, missingness percentage, data type, dataset size, computational resources, and purpose of the analysis.

VI. FACTORS AFFECTING METHOD SELECTION

The percentage of missing data is an important factor when selecting a treatment method. If only a small fraction of values is missing and the missingness is approximately random, simple methods may be sufficient.

As the proportion of missing data increases, deletion becomes increasingly costly because more observations are discarded. In such cases, imputation methods can help preserve available information.

The pattern of missingness is also important. Missingness concentrated in one variable differs from scattered missingness across multiple variables. Time dependent missingness may require methods that account for temporal relationships.

Data type must also be considered. Numerical, categorical, ordinal, and mixed-type datasets may require different approaches. For example, mean imputation is not appropriate for categorical variables.

The purpose of the analysis is another important factor. A simple method may be sufficient for exploratory analysis, whereas statistical inference may require an uncertainty-aware technique such as multiple imputation [4].

For machine-learning applications, the effect of imputation should be evaluated not only using imputation accuracy but also by examining the performance of the final predictive model [10].

VII. EVALUATION OF IMPUTATION METHODS

Evaluation of missing data methods is challenging because the actual values of naturally missing observations are usually unknown.

One common evaluation strategy is to begin with a complete or relatively complete dataset and artificially remove some observed values. The selected imputation technique is then applied, and the estimated values are compared with the original values.

For numerical variables, metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) can be used. For categorical variables, classification accuracy and related measures may be appropriate.

However, imputation accuracy alone is not sufficient. An imputation method may achieve low numerical error while changing the distribution or correlations between variables.

Therefore, evaluation should consider multiple aspects, including:

Imputation accuracy
Distribution preservation
Relationship preservation
Downstream machine-learning performance
Computational cost
Reproducibility
Uncertainty estimation

This broader evaluation provides a more reliable assessment of the usefulness of a missing-data technique [9], [10].

VIII. CHALLENGES AND LIMITATIONS

One of the major challenges is determining whether missingness is MCAR, MAR, or MNAR. These mechanisms are conceptual and cannot always be directly verified using observed data [1], [2].

MNAR is particularly difficult because the missing value itself may influence whether the value is observed. In such cases, standard imputation methods may produce biased results.

Scalability is another challenge. KNN requires distance calculations, multiple imputation generates several completed datasets, and iterative methods such as EM and random-forest imputation can become computationally expensive as dataset size increases [5], [8].

Deep-learning approaches can require significant computational resources and large datasets [7]. Although they may provide greater flexibility, their complexity may make them difficult to interpret.

Another important concern is bias amplification. If missingness is systematically higher for a particular group, an inappropriate imputation method may reinforce existing inequalities rather than correct them.

Finally, imputed values are estimates and should not automatically be treated as directly observed measurements. This is particularly important when the objective is statistical inference.

IX. RESEARCH GAPS AND FUTURE DIRECTIONS

The existing literature indicates several opportunities for future research.

First, adaptive missing-data systems could automatically analyse the missingness pattern and recommend an appropriate imputation method. Instead of applying one technique to every dataset, such systems could select methods according to dataset characteristics.

Second, future research can combine statistical modelling with machine learning. Hybrid approaches could provide the flexibility of machine-learning models while retaining uncertainty estimates associated with statistical methods.

Third, scalable missing-data techniques are needed for large datasets, streaming data, and distributed computing environments. Computational efficiency will become increasingly important as data volume continues to increase.

Fourth, fairness should receive greater attention. Future evaluation should examine whether imputation accuracy differs across groups and whether imputation changes downstream model performance unequally.

Finally, future studies should use realistic missingness simulations and evaluate both imputation quality and downstream predictive performance. Sensitivity analysis should also be included when the missingness mechanism is uncertain [10].

X. CONCLUSION

Missing data is an unavoidable characteristic of many real-world datasets, and the method used to handle it can significantly influence statistical analysis and machine-learning performance.

This review examined the major missing-data mechanisms, including MCAR, MAR, and MNAR, and discussed commonly used approaches such as deletion, statistical imputation, regression imputation, multiple imputation, EM, KNN, random-forest methods, and deep-learning approaches [1]– [10].

The comparison demonstrates that there is no universally best missing-data handling technique. Simple approaches remain useful when the amount of missingness is small and approximately random. More advanced approaches become valuable when relationships among variables, uncertainty, nonlinear patterns, or high-dimensional structures need to be preserved.

A reliable missing-data strategy should therefore begin by understanding the reason and pattern of missingness. The selection of a method should also consider data type, missingness percentage, computational resources, and the purpose of the analysis.

Future research should focus on adaptive, uncertainty-aware, scalable, and fairness-conscious approaches. Such developments can improve the reliability of data-driven systems and reduce the possibility that preprocessing decisions introduce hidden bias or misleading conclusions.

REFERENCES

[1] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[2] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[3] S. Devikar, M. Hinge, and S. S. Shevkari, “Human vs. machine: A review of AI’s impact on language learning interaction or replacing human interaction,” International Journal for Multidisciplinary Research, vol. 7, no. 3, 2025, doi: 10.36948/ijfmr. 2025.v07i03.49459.

[4] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in R,” Journal of Statistical Software, vol. 45, no. 3, pp. 1–67, 2011.

[5] D. J. Stekhoven and P. Bühlmann, “MissForest—non-parametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.

[6] O. Troyanskaya et al., “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[7] J. Yoon, J. Jordon, and M. van der Schaar, “GAIN: Missing data imputation using generative adversarial nets,” in Proc. 35th International Conference on Machine Learning, 2018, pp. 5689–5698.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society: Series B, vol. 39, no. 1, pp. 1–38, 1977.

[9] H. Chinni, Y. K. Sharma, S. Shevkari, J. L. Vydehi, Saurabh, and M. Kavitha, “Comparative evaluation of custom CNN architectures and transfer learning models for image classification in PyTorch,” in Proc. 2026 13th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2026, pp. 1–6, doi: 10.23919/INDIACom70271.2026.11525435.

[10] T. Emmanuel *et al.*, “A survey on missing data in machine learning,” *Journal of Big Data*, vol. 8, Art. no. 140, 2021.