Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(6), 81-102
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Missing Data Handling: A Comprehensive Review, Taxonomy, and Comparative Evaluation

Ikram Chourib

Paris, France
Email: chourib.ikram@gmail.com

How to cite this paper: Chourib, I. (2025) Missing Data Handling: A Comprehensive Review, Taxonomy, and Comparative Evaluation. *Journal of Computer and Communications*, **13**, 81-102.
https://doi.org/10.4236/jcc.2025.136006

Received: May 14, 2025
Accepted: June 16, 2025
Published: June 19, 2025

Abstract

Missing data remains a persistent and pervasive challenge across a wide range of domains, significantly impacting data analysis pipelines, predictive modeling outcomes, and the reliability of decision-making processes. This paper presents a comprehensive and updated review of missing data handling techniques that entail both traditional statistical methods and state-of-the-art graph-based and machine-learning approaches. A novel taxonomy is introduced, classifying strategies into three principal categories: preprocessing techniques, graph-based imputations, and algorithms inherently tolerant to missing values. Particular emphasis is placed on recent advancements in deep learning architectures, including Generative Adversarial Imputation Networks (GAIN), Self-Attention Imputation for Time Series (SAITS), and MissFormer, as well as graph-based methods such as Graph Recovery Imputation Network (GRIN) and Temporal Spatial Imputation Graph Neural Network (TSI-GNN). These models demonstrate notable improvements in handling complex missingness patterns and scaling to large heterogeneous datasets. To complement the theoretical review, an empirical evaluation was conducted on two benchmark datasets (Heart Disease and Kidney Disease), examining the effectiveness and limitations of various imputation strategies under different missingness scenarios. The results underscore the critical importance of adapting missing data handling techniques to the nature of the dataset, the underlying missingness mechanism, and the proportion of missing entries. Finally, the paper outlines promising research directions, advocating for the development of lightweight, explainable, and scalable models; online adaptive imputation strategies for streaming data; multimodal data integration techniques; and privacy-preserving imputation frameworks within federated and decentralized learning environments. Addressing these challenges is essential for building the next generation of reliable, transparent, and intelligent data-driven sys-

tems.

Keywords

Missing Data, Data Imputation, Deep Learning, Machine Learning

1. Introduction

Missing data is a widespread issue in a wide range of real-world applications, including healthcare, finance, environmental monitoring, and social sciences. The presence of incomplete information can severely compromise subsequent data analysis and decision-making processes, often resulting in biased conclusions and reduced model performance [1]-[3]. Generally, the term "missing" refers to the absence of an expected feature value due to various causes, such as data loss, non-response, recording errors, or inapplicability to a specific case. For example, in e-health applications, missing values occur frequently when patients do not complete all required fields or when technical errors arise during data collection [3].

One of the earliest strategies for addressing missing data involved the deletion of incomplete records. While such deletion-based approaches may be acceptable for datasets with redundant or non-critical information, they are generally discouraged in domains where data completeness is vital—such as medicine, finance, or risk assessment—as they can lead to significant information loss, sampling bias, and reduced statistical power. Consequently, foundational work by Little and Rubin [4] and the Working Group on Statistical Inference [5] advocated for more robust methodologies, including single and multiple imputation, as well as model-based approaches such as Expectation-Maximization (EM).

Over the past two decades, the field has seen the emergence of broader classifications. Notably, García-Laencina *et al.* [6] proposed a taxonomy encompassing record deletion, classical imputations, model-based approaches, and machine-learning methods. However, these earlier taxonomies often fail to account for two important methodological advances: 1) the rise of deep learning-based imputations capable of modeling high-dimensional and non-linear data structures, and 2) the development of Graph Neural Network (GNN)-based methods that exploit relational and topological information for imputation. Moreover, prior frameworks rarely distinguish algorithms that are intrinsically robust to missing data, *i.e.* those capable of learning from incomplete inputs without any preprocessing or imputation stage.

To address these gaps and offer a more comprehensive framework aligned with recent methodological developments, this study introduces a novel taxonomy that categorizes missing data handling techniques into three principal groups (as illustrated in **Figure 1**):

- Preprocessing Approaches: This category includes both classical and modern techniques applied before model training. It encompasses deletion-based strat-

egies (e.g. listwise and casewise deletion), simple statistical imputations (mean, median, mode), model-based imputations (e.g. regression, KNN, EM), and multiple imputation methods (e.g. MICE, MIWAE). It also integrates recent deep learning-based techniques such as GAIN, BRITS, SAITS, and MissFormer, as well as fuzzy clustering-based imputations.

- Graph-Based Imputation Approaches: These emerging methods utilize graph neural networks (e.g. GRIN, TSI-GNN) to model complex dependencies and relationships between data entities, offering a structured alternative particularly suited for relational, temporal, or spatially correlated datasets.
- Algorithms Intrinsically Supporting Missing Data: This group comprises learning algorithms that natively handle missing inputs during training, eliminating the need for prior imputation. Examples include decision trees (e.g. J48, C4.5), Random Forests, SMO-based SVMs, and advanced ensemble methods like XGBoost, LightGBM, and CatBoost.

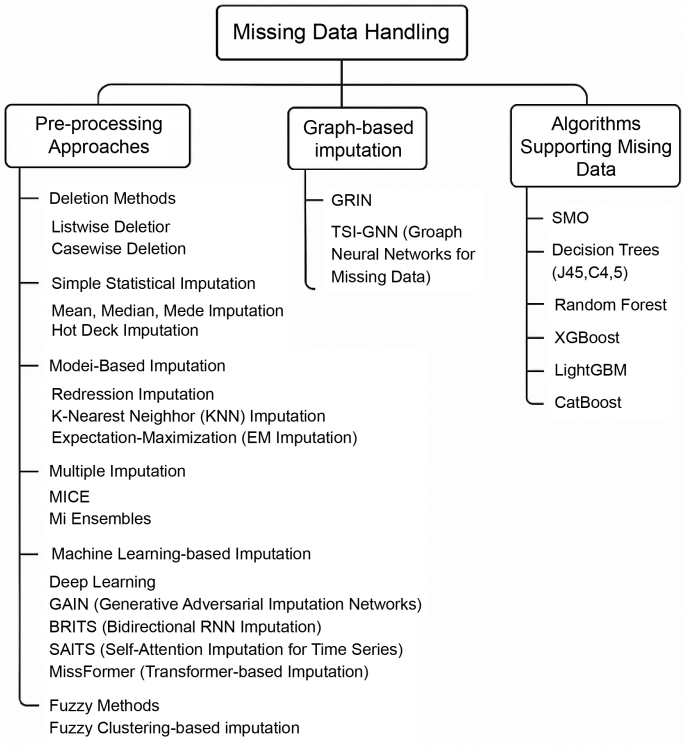

Figure 1. Proposed taxonomy of missing data handling methods, categorized into pre-processing approaches, graph-based imputation techniques, and algorithms intrinsically supporting missing values.

Unlike previous classifications, the proposed taxonomy explicitly integrates modern algorithmic paradigms and emphasizes the alignment between imputation strategies and the structural characteristics of the data. It reflects the progression from traditional statistical methods toward more scalable, adaptive, and context-aware learning frameworks. To clarify the added value of this framework, **Figure 1** presents a comparative synthesis with existing taxonomies, highlighting the distinc-

tive contributions and enhanced comprehensiveness of the present classification in light of current methodological trends.

The remainder of this paper is organized as follows. Section 2 presents the theoretical foundations of missing data, including the principal patterns and mechanisms underlying missingness. Section 3 provides a comprehensive review and taxonomy of established missing data handling techniques. Section 4 describes the empirical comparative study conducted on the Heart Disease and Kidney Disease datasets, followed by a critical analysis of the results. Finally, Section 5 concludes the paper and discusses promising directions for future research.

2. Theoretical Foundations of Missing Data

In the knowledge discovery process, data preparation constitutes one of the most critical and time-consuming phases, with a direct impact on the reliability and validity of research outcomes. In particular, variable selection involves identifying a relevant subset of predictors from a broader set of attributes. A common practice is to eliminate variables with a substantial proportion of missing values (e.g. exceeding 50%), which is generally advisable but not without risk [7]. Removing variables may lead to a loss of predictive power, a reduction in the ability to detect statistically significant associations, and potential introduction of bias, ultimately affecting the robustness of the analysis. Consequently, feature selection should be carefully adapted to the pattern and mechanism of missing data, with imputation potentially applied either before or after the variable selection step [7].

Addressing missing data typically involves the following general steps:

- Identifying the causes of missing data;
- Analyzing the patterns of missing data;
- Investigating the underlying missing data mechanism;
- Selecting and applying the appropriate imputation method.

In this section, each of these steps is discussed in detail.

2.1. Causes of Missing Data

Researchers frequently encounter missing data challenges, particularly during the data collection phase. According to the literature [8]-[10], the causes of missing data can generally be classified into two categories: intentional and unintentional.

Intentional Missing Data

During data collection, specific responses or conditions may render the collection of certain features unnecessary. For instance, the answer to a preliminary question might make subsequent questions irrelevant for some respondents. In such cases, missingness occurs by design, reflecting the logical structure of the data acquisition process.

Unintentional Missing Data

This category encompasses missingness arising from real-world challenges such as errors in manual data entry, sensor failures, system malfunctions, or difficulties in accessing certain measurements. Unintentional missingness is often random

and more problematic, as it can compromise data quality if not properly addressed.

2.2. Missing Data Patterns

A pattern refers to a vector, case, or observation characterized by n features, each supporting continuous, discrete, or symbolic values. Mathematically, a pattern can be represented as:

$$X = [x_1, x_2, \cdots, x_n]$$

Figure 2 illustrates two examples: (a) a complete pattern without missing values and (b) a pattern containing missing data, where black cells represent missing entries [6] [11].

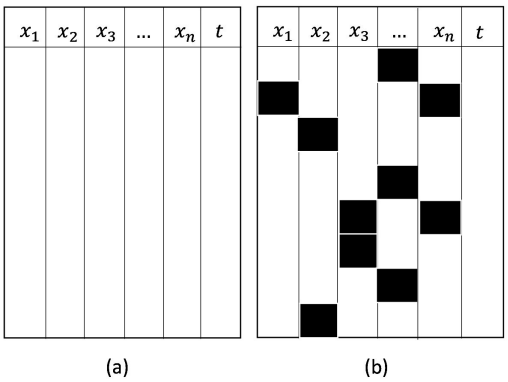

Figure 2. (a) A pattern without missing data; (b) A pattern with missing data.

Missing data patterns can vary considerably. Figure 3 shows six prototypical missing data structures commonly encountered, following the notation from [12].

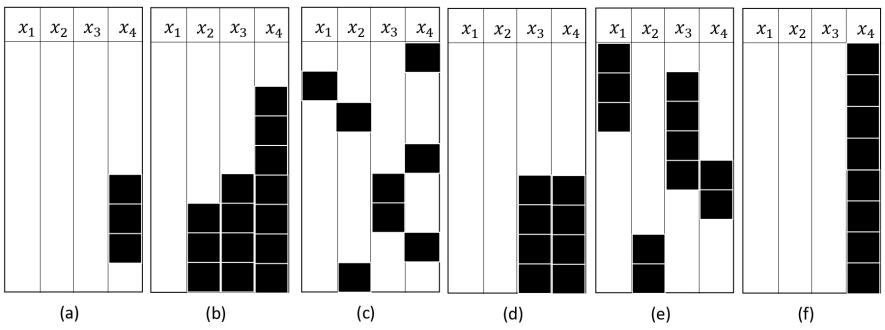

Figure 3. Prototypical missing data patterns.

The six principal patterns are described as follows:

- Univariate Pattern (Figure 3(a)): Missingness occurs in a single feature (e.g. only contains missing values).
- Monotone Pattern (Figure 3(b)): Missing data follows a staircase structure; once a value is missing, subsequent entries are also missing.

- General (Arbitrary) Pattern (**Figure 3(c)**): Missing data is scattered randomly throughout the dataset.
- Unit Non-Response (**Figure 3(d)**): Entire rows are missing, typically due to complete participant dropout.
- Planned Missingness (**Figure 3(e)**): Different subsets of features are intentionally missing across cases to reduce respondent burden while maximizing data coverage.
- Latent Variable Missingness (**Figure 3(f)**): An entire feature is missing across all instances, often due to unmeasured constructs [13].

2.3. Missing Data Mechanisms

Missing data mechanisms describe the relationship between missingness and the observed or unobserved values in the dataset. Following the seminal work of Little and Rubin [4], three principal mechanisms are distinguished:

Missing Completely at Random (MCAR)

The probability of a value being missing is independent of both observed and unobserved data. MCAR typically occurs in datasets with less than 5% missingness [4] and implies that the observed cases are representative of the full sample. Formally:

$$P(\text{Missing} \mid \text{Complete Data}) = P(\text{Missing}) \tag{1}$$

Missing at Random (MAR)

The probability of missingness depends only on the observed data but not on the unobserved (missing) values themselves. Under MAR, correct modeling of the observed data can mitigate missingness bias. Formally:

$$P(\text{Missing} \mid \text{Complete Data}) = P(\text{Missing} \mid \text{Observed Data}) \tag{2}$$

Not Missing at Random (NMAR)

The probability of missingness depends on the unobserved (missing) data itself. NMAR poses the greatest challenge because traditional imputation methods generally assume MCAR or MAR. Formally:

$$P(\text{Missing} \mid \text{Complete Data}) = P(\text{Missing} \mid \text{Observed Data}, \text{Missing Data}) \tag{3}$$

Among these mechanisms, NMAR is particularly problematic and often requires specialized modeling approaches. Ignoring the mechanism underlying missingness can lead to biased estimates and compromised inference. In particular, listwise deletion under MCAR remains unbiased but sacrifices efficiency, while under MAR or NMAR, it introduces systematic biases.

3. Review of Missing Data Handling Techniques

The development of missing data handling techniques has been motivated by the increasing complexity, volume, and heterogeneity of datasets in modern applications. These techniques can be broadly categorized into four main groups: 1) simple statistical imputation methods, 2) multiple imputation strategies, 3) model-based ap-

proaches, and 4) machine learning algorithms natively tolerant to missing data. A detailed discussion of each category is provided below.

3.1. Simple Statistical Imputation

Among the most intuitive strategies, mean imputation replaces missing entries with the mean of the corresponding variable. Its simplicity and computational efficiency make it attractive, particularly for preliminary analyses or low-stakes applications [14]. However, the method has serious drawbacks: it artificially reduces data variance, weakens feature correlations, and may bias estimates, especially under non-MCAR missingness mechanisms. Zhang [15] emphasized that mean, median, and frequent-value imputations introduce systematic errors in variance and covariance estimation, leading to unreliable statistical inferences. Farhangfar *et al.* [16] confirmed that simple imputations should be restricted to cases where missingness is rare (less than 10%) and where model robustness to data quality is not a critical requirement. Despite its limitations, mean imputation remains widely used in industrial settings where interpretability and speed outweigh concerns about statistical rigor.

3.2. Multiple Imputation (MI)

Introduced by Rubin *et al.* [4], Multiple Imputation (MI) has become a gold standard for addressing missing data, especially under MAR assumptions. MI accounts for uncertainty by creating multiple complete datasets, analyzing each separately, and combining the results according to Rubin's rules [17]. This approach preserves variability and improves inferential validity compared to single imputation.

Recent advances include:

- MICEs (Multivariate Imputation by Chained Equations), enabling variable-by-variable conditional modeling.
- Deep Multiple Imputation techniques, such as MI with Deep Ensembles [18] and MIWAE [19], which use deep generative models to better capture complex feature distributions.

While MI is computationally demanding, its ability to model the imputation uncertainty makes it preferable for high-stakes analyses, such as medical research and policy evaluations.

3.3. Regression Imputation

Regression imputation imputes missing values based on predictive models built from observed data [3]. Linear or logistic regression models are typically used to estimate missing entries, preserving inter-variable dependencies. However, this method is not without challenges:

- If the underlying model is misspecified, regression imputation can lead to biased predictions [16].
- Variability is often underestimated unless random residuals are added to the predictions [15].

To address these limitations, Bayesian Regression Imputation (BRI) [20] models the full posterior distribution of missing values, thereby incorporating uncertainty into the imputation process. Regression-based methods remain powerful when feature relationships are strong and correctly specified but require careful model validation.

3.4. K-Nearest Neighbors (KNN) Imputation

The K-Nearest Neighbors (KNNs) approach, originally proposed by Cover *et al.* [21], imputes missing values by leveraging local similarity among instances. It operates through: Measuring distances between observations, identifying K closest neighbors and averaging their feature values to impute missing data.

KNN imputation is non-parametric, meaning it does not assume any particular data distribution. However, its computational cost grows with dataset size [22]. Recent developments, such as Approximate KNN using Locality-Sensitive Hashing (LSH) [23], significantly reduce the computational burden, making KNN viable even for larger datasets.

KNN remains particularly attractive for imputation in biomedical and environmental datasets where local correlations dominate.

3.5. Expectation-Maximization (EM) Imputation

The Expectation-Maximization (EM) algorithm [24] provides a statistically grounded approach to imputation under MCAR or MAR assumptions. It iteratively alternates between:

- E-step: Estimating missing values given the current model parameters.
- M-step: Updating parameters based on the imputed data.

Convergence is declared once changes between iterations fall below a threshold. While EM offers elegant theoretical properties, its practical limitations include:

- Slow convergence, especially in high-dimensional settings.
- Sensitivity to initial parameter values.

Stochastic versions (SEM) and Variational EM (VEM) [25] introduce approximations to enhance convergence speed. EM remains a reference method for structured missingness problems, particularly in psychometrics and clinical studies.

3.6. Multi-Layer Perceptron (MLP) Imputation

Multi-Layer Perceptrons (MLPs) are neural network models capable of modeling complex non-linear relationships, making them well-suited for imputing missing data [6]. Theoretical underpinnings of MLP-based imputation lie in their universal approximation capabilities, enabling them to effectively capture intricate data structures and interdependencies. However, the performance and applicability of MLPs heavily depend on the dataset size and complexity of missingness patterns. Training separate networks for distinct missingness configurations significantly increases computational complexity, demanding considerable computational resources and careful hyperparameter tuning. Furthermore, small datasets exacer-

bate the risk of overfitting, necessitating regularization techniques like dropout, L2 regularization, and early stopping. Empirical successes have been documented in domains such as breast cancer diagnosis [26], where data complexity justified the computational overhead, and environmental monitoring [27], where the non-linear relationships between variables benefited from MLPs' modeling strength. The advancement of self-supervised learning frameworks is anticipated to mitigate existing limitations, further enhancing MLP-based imputation effectiveness in diverse scenarios.

3.7. Algorithms Natively Supporting Missing Data

Certain machine learning algorithms inherently accommodate missing data without prior imputation by leveraging algorithm-specific mechanisms. Decision trees (e.g. J48, C4.5) use surrogate splits to manage missing features, theoretically allowing robust predictions despite incomplete inputs. Naïve Bayes models handle missingness through marginal probability estimations, assuming conditional independence among features, which simplifies missing data treatment. Random forests inherently accommodate missing data during bootstrapping by selecting optimal splits that minimize prediction error despite incomplete observations [28]. Support Vector Machines (SVMs), particularly Sequential Minimal Optimization (SMO) variants, exhibit robustness by optimizing decision boundaries based solely on observed dimensions [29]. Nonetheless, these methods have practical limitations; for instance, extensive missingness in critical features can reduce predictive accuracy, and surrogate splits in decision trees can sometimes obscure model interpretability. Such algorithms remain particularly advantageous for rapid prototyping and real-time decision-making, where preprocessing and imputation are impractical or computationally prohibitive.

3.8. Deep Learning-Based Imputation

Recent advancements in deep learning have significantly expanded capabilities in missing data imputation, especially for high-dimensional and complex datasets. Methods such as Generative Adversarial Imputation Networks (GAINs) [30] leverage adversarial training to generate realistic imputations, theoretically benefiting from the generative modeling of data distributions. BRITS [31] employs recurrent neural networks to effectively capture temporal dependencies, ideal for time-series datasets with sequential missingness patterns. Transformer-based approaches, including SAITS [32] and MissFormer [33], utilize self-attention mechanisms capable of modeling extensive long-range dependencies within sequences or structured tabular data. Despite these strengths, deep learning methods inherently require extensive computational resources, careful hyperparameter tuning, and robust training frameworks to prevent model instability or mode collapse, especially in adversarial contexts. Empirical comparisons suggest SAITS and BRITS outperform traditional imputation methods in sequential contexts, while MissFormer demonstrates considerable potential in general tabular datasets due to its direct

handling of heterogeneous missing data types.

3.9. Graph Neural Networks (GNNs) for Missing Data

Graph Neural Networks (GNNs) exploit relational structures inherent in many datasets to improve imputation performance. Theoretical foundations for graph-based imputation methods, such as GRIN [34] and TSI-GNN [35], involve leveraging node relationships and neighborhood information to infer missing values effectively. By modeling spatiotemporal dynamics and relational contexts explicitly, GNNs provide superior performance in structured domains such as social networks, sensor networks, and biological systems. However, significant limitations include the necessity for well-defined relational data, the computational cost associated with constructing and training on large graphs, and the sensitivity of GNN models to noise and incorrect relational assumptions. Practical implementations should carefully consider the dataset's suitability, particularly evaluating relational structure strength, node interdependencies, and scalability concerns.

3.10. Tree-Based Algorithms with Native Missing Data Handling

Modern ensemble tree algorithms, including XGBoost [36], LightGBM [37], and CatBoost [38], natively manage missing data effectively through built-in mechanisms such as optimal split-direction learning and missing value masking. Theoretically, these algorithms mitigate missingness by systematically identifying splits that minimize prediction error, thus maintaining robustness without explicit data imputation. Although effective, these methods can reduce model interpretability due to their implicit handling of missing values, and high levels of missingness might sometimes decrease model performance compared to explicit imputation strategies. Comparative evaluations indicate these methods are especially beneficial in large-scale, sparse datasets, offering computational efficiency and robust performance when preprocessing and explicit imputation are impractical. Clearly articulating these scenarios facilitates better methodological choices tailored to specific application contexts.

3.11. Comparative Analysis and Critical Discussion

Handling missing data remains a critical challenge in contemporary data analysis pipelines across various scientific and industrial domains. The primary objective of this study is to present a comprehensive review of classical and modern techniques for recovering missing data, emphasizing machine learning-based and data-driven methodologies. The choice and effectiveness of imputation methods fundamentally depend on the specific patterns and mechanisms underlying the missing data. Structured patterns, such as monotone or planned missingness, naturally support targeted imputation strategies like regression-based methods or Expectation-Maximization (EM), benefiting from clear and predictable feature dependencies. Conversely, arbitrary or general patterns characterized by random and irregular missing entries necessitate robust and flexible approaches, including Random Forests

or advanced deep-learning models (e.g. GAIN, MissFormer).

Additionally, understanding the missing data mechanisms—Missing Completely at Random (MCAR), Missing at Random (MAR), or Not Missing at Random (NMAR)—is essential for method selection. Under MCAR conditions, simpler methods such as mean imputation or listwise deletion remain unbiased yet may sacrifice statistical efficiency. For MAR data, techniques that leverage observed correlations, such as Multiple Imputation by Chained Equations (MICEs), provide unbiased and efficient imputations. NMAR scenarios, representing the most challenging conditions, necessitate specialized modeling frameworks such as selection or pattern-mixture models, explicitly addressing dependencies related to the unobserved values. Thus, accurately identifying and characterizing missing data patterns and mechanisms is crucial for selecting appropriate imputation methods, ultimately ensuring their effectiveness and validity.

The success of missing data recovery strategies is closely tied to the initial feature selection process. Selecting informative and relevant features not only mitigates the adverse impact of missingness but also enables the effective application of algorithms that natively tolerate missing values. In contrast, retaining non-informative or sparsely populated features can degrade model performance and lead to biased inferences, particularly in high-dimensional settings. Feature selection thus must be adapted to the missing data patterns and mechanisms at play, reinforcing its strategic importance in the imputation pipeline. Simple statistical imputation methods—such as mean, median, or mode imputation—continue to be widely used due to their simplicity and computational efficiency. However, as confirmed in this study and supported by the literature, these methods tend to artificially reduce data variability and neglect inter-feature dependencies. Consequently, they often introduce biases and degrade model reliability, especially when the missingness rate exceeds 10% - 20%. Advanced statistical approaches, such as Expectation-Maximization (EM) and Multiple Imputation (MI), provide better estimates under the Missing at Random (MAR) assumption by modeling the underlying data distribution. Nevertheless, these methods are computationally intensive and show limitations when faced with high-dimensional datasets or Not Missing at Random (NMAR) conditions, which are increasingly encountered in real-world data.

Recent advances in machine learning have significantly enhanced the capacity to handle missing data. Techniques based on Generative Adversarial Networks (e.g. GAIN), Variational Autoencoders, Transformer-based architectures (e.g. SAITS, MissFormer), and Graph Neural Networks (e.g. GRIN, TSI-GNN) have demonstrated superior performance compared to classical methods. These models are particularly adept at:

- Capturing complex, non-linear dependencies between features.
- Adapting to heterogeneous and structured missingness patterns.
- Scaling to large datasets through GPU-parallelized computation.

Such capabilities make deep learning-based methods highly suitable for critical application domains such as healthcare analytics, financial risk prediction, and

Internet of Things (IoT) monitoring.

In fuzzy systems, missing inputs exacerbate uncertainty propagation, thereby impairing the reasoning process and decision reliability. Hybrid strategies that combine fuzzy logic with machine learning-based imputation or fuzzy clustering show promise in overcoming these limitations, especially in sensitive fields like medical diagnosis and environmental monitoring.

The selection of a missing data handling technique should be driven by multiple factors, including:

- The missingness mechanism (MCAR, MAR, NMAR).
- The dimensionality and size of the dataset.
- The criticality of the target application.
- Computational resource availability.

Traditional statistical methods remain appropriate for small, low-dimensional datasets with moderate missingness rates. Conversely, modern deep learning and graph-based models offer scalable and robust alternatives for complex, large-scale data environments.

To provide a consolidated view, Table 1 presents a comparative summary highlighting the key distinctions between classical statistical imputation methods and modern deep learning-based techniques, encompassing their respective strengths, limitations, and typical application domains.

Table 1. Comparative analysis of traditional and modern imputation techniques for missing data.

Criteria	Classical Methods (Mean, EM, MICE, KNN, Regression)	Modern Methods (GAIN, MissFormer, GRIN, TSI-GNN)
Ease of Implementation	Very simple and fast to deploy	More complex architectures requiring careful design
Computational Time	Low to moderate (depending on method)	Moderate to high (requires GPU acceleration)
Imputation Quality	Good for low missingness rates (≤10%)	Excellent, even for high missingness rates
Scalability to Large Datasets	Limited (e.g. EM and MICE scale poorly)	Very good (deep models scale efficiently with parallelization)
Handling of Complex Data Structures	Limited capacity	Strong ability to capture non-linear and long-range dependencies
Robustness to MCAR/MAR/NMAR	Mainly effective for MCAR and MAR	Better suited for MAR and partially NMAR scenarios
Interpretability	High (transparent and explainable)	Lower (black-box nature of deep learning models)
Typical Applications	Small datasets, traditional research settings	Big data analytics, IoT, healthcare, financial forecasting

4. Experimental Setup

4.1. Datasets

Two datasets were employed to evaluate the performance of various missing data handling techniques:

Heart Disease Dataset: Sourced from https://www.kaggle.com/c/heart-disease-uci, this dataset comprises 303 instances and 14 features. It is fully complete, containing no missing values (0% missing-

ness).

Kidney Disease Dataset: Also obtained from https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset, this dataset consists of 400 instances and 26 features. It is an inherently incomplete dataset, containing real missing values. The dataset includes two diagnostic classes: "sick" (59%) and "not sick" (41%).

All datasets were preprocessed through normalization prior to applying missing data techniques.

4.2. Simulation of Missing Data

To systematically evaluate imputation methods, artificial missingness was introduced into the complete heart disease dataset at varying levels: 10%, 20%, 30%, 40%, and 50%. Missing values were inserted completely at random (MCAR mechanism) across the dataset to simulate realistic scenarios. This procedure generated five experimental settings, each corresponding to a different proportion of missing data.

4.3. Imputation Methods

For the heart disease dataset, five different imputation strategies were applied to handle missing data: Mean Imputation, K-Nearest Neighbors (KNNs) Imputation, Multiple Imputation by Chained Equations (MICEs), Regression Imputation, Expectation-Maximization (EM) Imputation.

For the KNN imputation method, the optimal number of neighbors (k) was determined by minimizing the Mean Error Rate (MER) for each missingness scenario.

For the kidney disease dataset, the following classifiers were used to assess the impact of missing data handling: J48 Decision Tree, Naive Bayes Classifier, Random Forest, REP Tree, Support Vector Machine (via Sequential Minimal Optimization, SMO) and the XGBoost (Extreme Gradient Boosting) algorithm was included in the experimental evaluation. XGBoost is a modern ensemble method that natively supports missing values by learning optimal split directions during training, eliminating the need for explicit imputation. Its high scalability and robustness make it a strong candidate for structured data classification tasks.

Each model was trained and evaluated after applying appropriate missing data-handling techniques.

4.4. Evaluation Metrics

Performance evaluation was conducted using the following criteria:

- Root Mean Square Error (RMSE): was calculated to quantify the average magnitude of imputation errors, as commonly recommended in imputation evaluation studies [39].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

- Receiver Operating Characteristic (ROC) Analysis: For classification tasks, ROC analysis was used to assess the diagnostic ability of classifiers across varying discrimination thresholds. Specifically, the Area Under the ROC Curve (AUC) was calculated, where an AUC of 1.0 indicates perfect classification, and an AUC of 0.5 reflects random guessing [40].
- Kappa Statistic: To assess the reliability of the classification results beyond chance agreement, the kappa coefficient was calculated [41]. Kappa values closer to 1 indicate strong agreement between predictions and actual class labels, while values near 0 reflect agreement no better than random chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

4.5. Results

4.5.1. Effect of K-Value of Imputation Performance

To determine the optimal number of neighbors K for KNN-based imputation, an empirical evaluation was conducted across varying levels of missingness (10% - 50%) in the Heart disease dataset. Figure 4 illustrates the evolution of the error rate as a function of K for each missingness level. As expected, the error rate generally decreases with increasing K, although fluctuations are observed due to dataset variance. Notably, the lowest error rate is achieved for K = 16 when 10% of the data is missing, and for K = 7 when the missingness reaches 50%. This analysis highlights the importance of tuning K for optimal performance in KNN imputation.

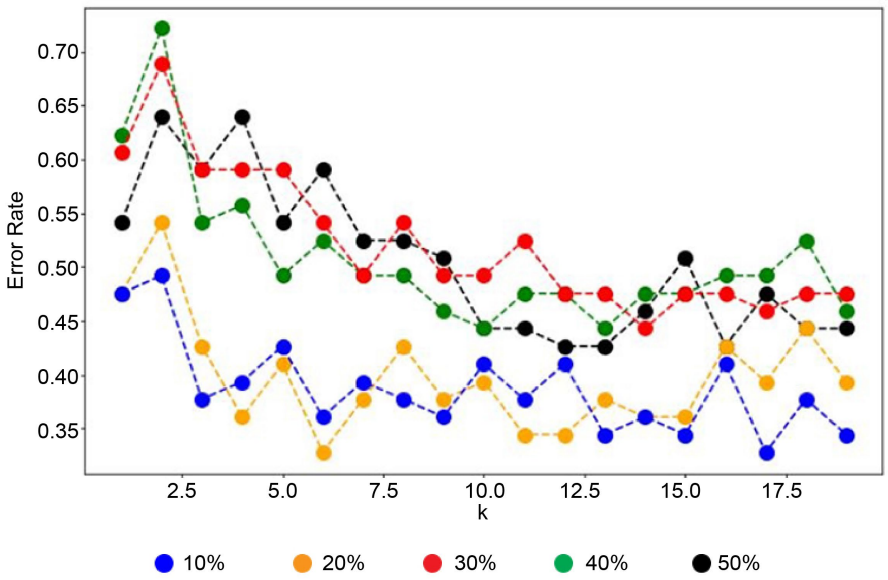

Figure 4. Mean error rate across different values of K for five levels of randomly introduced missingness in the heart disease dataset.

4.5.2. RMSE Evaluation of the Heart Disease Dataset

Table 2 presents the Root Mean Square Error (RMSE) values for five imputation methods (mean, KNN, MICE, regression, and EM) applied to the heart disease dataset under five missing data scenarios. The Expectation-Maximization (EM) method consistently achieved the lowest RMSE across all levels of missingness, confirming its robustness in preserving data integrity. In contrast, regression-based imputation exhibited the poorest performance, particularly at higher missingness levels (e.g. RMSE = 0.67 at 50%). These findings indicate that more sophisticated probabilistic or iterative methods are better suited for handling significant data gaps.

Table 2. RMSE of imputation methods under varying missing data scenarios in the heart disease dataset.

Scenario	Mean	KNN	MICE	Regression	EM
10%	0.19	0.25	0.22	0.32	**0.16**
20%	0.22	0.30	0.25	0.39	**0.20**
30%	0.24	0.40	0.31	0.40	**0.21**
40%	0.29	0.45	0.35	0.60	**0.31**
50%	0.33	0.50	0.40	0.67	**0.35**

4.5.3. Classification Performance of the Kidney Disease Dataset

The classification performance of five algorithms (J48, Naïve Bayes, Random Forest, REP Tree, SMO and XGBoost) was assessed using the kappa statistic before and after applying mean imputation. As shown in Table 3, mean imputation substantially improved agreement scores across all classifiers. The SMO and XGBoost algorithm achieved the highest kappa value after imputation ($\kappa = 0.88$, $K = 0.9$), indicating strong predictive consistency.

Table 3. Kappa statistics for classification algorithms on the kidney disease dataset, with and without imputation.

Method	J48	Naïve Bayes	Random Forest	REP Tree	SMO	XGBoost
Without Imputation	0.65	0.66	0.67	0.48	0.70	0.73
Mean Imputation	0.85	0.83	0.82	0.49	**0.88**	**0.9**

4.5.4. ROC Curve Evaluation of the Kidney Disease Dataset

To further assess classification performance, Receiver Operating Characteristic (ROC) analysis was conducted. Table 4 shows the Area Under the ROC Curve for the six classifiers, with and without imputation. Across all models, performance improved after imputation. SMO and XGBoost again outperformed other methods, achieving an ROC higher than 95%. The lowest performance was observed for REP Tree, which showed limited sensitivity to imputation.

Table 4. ROC AUC scores for classifiers on the kidney disease dataset, with and without imputation.

Method	J48	Naïve Bayes	Random Forest	REP Tree	SMO	XGBoost
Without Imputation	90.1%	82%	83.3%	48.9%	93%	93.4%
With Imputation	94%	85%	85.2%	50.1%	**96.5%**	**97.2%**

4.6. Discussion

The experimental analysis provided quantitative insights into how different missing data handling strategies influence downstream tasks such as imputation accuracy and classification performance across datasets of varying complexity and size. In this section, we systematically discuss the empirical trends, emphasizing comparative granularities among similar methodological approaches.

Results obtained from the Heart Disease dataset (Table 2) highlight that the Expectation-Maximization (EM) approach consistently outperformed alternative imputation methods-including mean, KNN, MICE, and regression-based approaches-especially at lower levels of missingness (below 40%). EM yielded the lowest RMSE values, reflecting its capability to effectively model underlying data distributions. For example, at missingness levels of 10% and 20%, EM achieved RMSE scores of 0.16 and 0.20, respectively, demonstrating notable superiority. However, as missingness increased beyond 40%, the performance advantage of EM significantly diminished, allowing simpler imputation methods such as mean imputation to become competitively viable. This shift underscores how substantial missingness erodes data structures, thus diminishing the relative benefit of sophisticated probabilistic models.

Moreover, a detailed analysis of KNN hyperparameters (Figure 4) revealed that the optimal number of neighbors (K) is sensitive to the proportion of missing data, emphasizing the necessity of adaptive tuning for instance-based methods.

In classification experiments conducted on the Kidney Disease dataset, machine learning algorithms inherently robust to missing values, such as Random Forest, SMO, and XGBoost, demonstrated superior performance compared to models trained on datasets imputed with simpler techniques, such as mean imputation. Notably, both SMO and XGBoost achieved the highest Kappa scores and ROC AUC values (Table 3 and Table 4). Specifically, SMO achieved a Kappa score of 0.88 and ROC AUC of 96.5%, closely matched by XGBoost, which exhibited consistently robust performance across varying missingness scenarios. These results underscore the robustness of gradient boosting and kernel-based models in internally mitigating the detrimental effects of missing data without requiring explicit imputation procedures.

Crucially, while the results provide clear quantitative assessments of method performance, they also reveal a lack of detailed comparative insights among similar deep-learning architectures within the same scenarios. Future empirical analyses should aim to incorporate finer-grained comparisons between deep-learning

approaches (e.g. GAIN versus MissFormer), elucidating how architectural differences influence their performance in handling identical missing data situations. Such granularity will facilitate deeper methodological understanding and enable researchers to make more informed decisions tailored precisely to their analytical contexts.

A consistent observation across both evaluated datasets is that no single imputation or modeling strategy universally excels. EM demonstrates strong efficacy in moderate missingness scenarios but loses its advantage with increasing missingness. Conversely, robust models like SMO, Random Forest, and XGBoost offer stable performance irrespective of explicit imputation, although their exact benefits depend on data complexity and missingness patterns.

The experimental findings stress the necessity of context-sensitive method selection. For datasets exhibiting moderate missingness and structured inter-feature dependencies, combining EM-based imputation with kernel-based (SMO) or gradient boosting (XGBoost) classification strategies is recommended. In contrast, for high missingness or datasets lacking strong feature correlations, simpler or hybrid imputation approaches may suffice. Incorporating performance metrics such as RMSE, ROC AUC, and Kappa scores into decision-making processes will further enhance the principled selection of suitable methodologies.

Finally, to support method selection decisions, Table 5 summarizes the evaluated imputation and modeling strategies, detailing their strengths, limitations, and suitable application contexts. This comparative overview underscores the importance of aligning method choices with the statistical properties of the dataset, analytical objectives, and practical constraints specific to each application.

Table 5. Comparative summary of missing data handling methods.

Method	Strengths	Limitations	Recommended Context
Mean Imputation	Fast; simple to implement	Ignores feature relationships; biased under MAR/MNAR	Small datasets with low missingness
KNN Imputation	Preserves local structure; non-parametric	Sensitive to k; high computational cost	Medium-scale data with non-linear dependencies
MICE	Captures multivariate relations; flexible	Slower convergence; less robust at high missingness	Structured data with moderate missingness
Regression Imputation	Captures linear dependencies	Underperforms with non-linear data; variance underestimation	Low-dimensional, linearly dependent data
EM	Statistically rigorous; best RMSE under MCAR	Sensitive to initialization; computationally demanding	Moderate missingness in structured datasets
Random Forest	Natively handles missingness; robust	Performance varies with missingness pattern	High-dimensional classification tasks
SMO (SVM)	High resilience; strong Kappa and ROC scores	Less interpretable; kernel tuning required	Binary classification with partial data (e.g. healthcare)
XGBoost	Native missing data handling; scalable and robust	Requires careful tuning; complexity increases with depth	Large-scale structured datasets with heterogeneous missingness

This comparative summary highlights the importance of aligning method selection with the statistical properties of the data, the analytical objectives, and the operational constraints of the target application.

5. Conclusions

Handling missing data remains a critical and persistent challenge across various domains where data quality directly impacts analytical outcomes and decision-making reliability. This study presented a comprehensive and structured review of classical and contemporary missing data handling techniques, encompassing methods ranging from simple statistical imputations to cutting-edge deep learning and graph-based solutions.

Through the development of an updated taxonomy, the evolution of missing data recovery approaches was systematically highlighted, demonstrating the growing influence of machine learning, deep learning, and graph neural networks in addressing increasingly complex, large-scale, and heterogeneous datasets. Experimental evaluations on diverse benchmark datasets confirmed that the effectiveness of imputation methods is strongly dependent on several factors, including the missingness mechanism (MCAR, MAR, NMAR), the dimensionality of the data, and the proportion of missing values.

While traditional methods remain well-suited for small datasets with low missingness rates, modern approaches, particularly those based on deep learning and graph structures, have become indispensable for achieving robust, scalable, and high-fidelity imputations in contemporary data environments. Furthermore, algorithms inherently tolerant of missing data offer valuable alternatives in scenarios where imputation is either computationally infeasible or prone to introducing biases.

Looking ahead, the continued development of lightweight, explainable, adaptive, and privacy-preserving missing data handling techniques will be crucial to meet the emerging needs of dynamic, distributed, and resource-constrained environments. Advancing research in this domain remains essential to enhance the reliability, interpretability, and scalability of intelligent systems across critical applications such as healthcare, finance, autonomous systems, and beyond.

6. Future Directions

Despite substantial progress, several challenges in missing data handling persist and require further investigation to fully address the growing complexity and dynamism of real-world datasets.

Although deep learning-based methods such as GAIN, MissFormer, and SAITS have demonstrated remarkable capabilities, they remain computationally demanding and often lack interpretability. Future research should prioritize the design of lightweight and explainable imputation architectures capable of operating in real-time environments, especially for applications in mobile health monitoring, autonomous systems, and edge computing.

Most existing imputation methods assume static datasets, which limit their applicability in dynamic contexts such as Internet of Things (IoT) networks, financial markets, and time-evolving databases. Developing online, adaptive imputation strategies that update missing value estimates in real-time, possibly coupled with reinforcement learning techniques, represents a promising research avenue.

Missing data in high-dimensional and multimodal datasets, such as genomics, medical imaging, and multimodal sensor fusion, presents unique challenges due to intricate and non-linear feature interdependencies. Future approaches should leverage multi-view learning, manifold learning, and advanced dimensionality reduction methods to efficiently capture underlying structures and deliver accurate imputations.

With the increasing emphasis on data privacy, especially in healthcare and finance, developing privacy-preserving imputation techniques suitable for federated and distributed learning frameworks has become a critical need. Federated imputation methods that perform local imputations without centralizing sensitive data will ensure regulatory compliance while maintaining model performance and generalizability.

Combining the strengths of graph neural networks (e.g. GRIN, TSI-GNN) with probabilistic reasoning frameworks and uncertainty quantification holds strong potential for advancing missing data handling. Such hybrid models could enhance robustness, interpretability, and reliability, particularly in high-stakes applications such as autonomous vehicles, clinical decision support systems, and financial forecasting.

In conclusion, advancing missing data recovery techniques will require interdisciplinary efforts, integrating innovations from deep learning, probabilistic modeling, graph theory, online learning, and privacy-preserving computation. Meeting these challenges is essential to enable the next generation of reliable, scalable, and intelligent data-driven systems.

Funding

This research received no external funding.

Author's Contribution

Ikram Chourib conceptualized the study and wrote the manuscript, performed the experiments and data analysis and reviewed and edited the final version.

Data Availability Statement

The datasets analyzed during the current study are publicly available. Two datasets were used to evaluate the performance of various missing data-handling techniques:

- **Heart Disease Dataset:** Available on https://www.kaggle.com/c/heart-disease-uci.
- **Kidney Disease Dataset:** available on https://www.kaggle.com/akshayksingh/kidney-disease-dataset.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Zor, K., Çelik, Ö., Timur, O., Yildirim, H. B. and Teke, A. (2018) Simple Approaches to Missing Data for Energy Forecasting Applications. *Proceedings of the* 16*th International Conference on Clean Energy*, Gazimagusa, 9-11 May 2018, 9-11.

[2] Laña, I., Olabarrieta, I., Vélez, M. and Del Ser, J. (2018) On the Imputation of Missing Data for Road Traffic Forecasting: New Insights and Novel Techniques. *Transportation Research Part C*: *Emerging Technologies*, **90**, 18-33. https://doi.org/10.1016/j.trc.2018.02.021

[3] Blankers, M., Koeter, M.W.J. and Schippers, G.M. (2010) Missing Data Approaches in Ehealth Research: Simulation Study and a Tutorial for Nonmathematically Inclined Researchers. *Journal of Medical Internet Research*, **12**, e1448. https://doi.org/10.2196/jmir.1448

[4] Little, R. and Rubin, D. (2019) Statistical Analysis with Missing Data, 3rd Edition, Wiley. https://doi.org/10.1002/9781119482260

[5] Wilkinson, L. (1999) Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, **54**, 594-604. https://doi.org/10.1037/0003-066x.54.8.594

[6] García-Laencina, P.J., Sancho-Gómez, J. and Figueiras-Vidal, A.R. (2010) Pattern Classification with Missing Data: A Review. *Neural Computing and Applications*, **19**, 263-282. https://doi.org/10.1007/s00521-009-0295-6

[7] Chourib, I., Guillard, G., Farah, I.R. and Solaiman, B. (2022) Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers. *IRBM*, **43**, 678-686. https://doi.org/10.1016/j.irbm.2022.02.002

[8] Peterkova, A., Nemeth, M. and Bohm, A. (2018) Computing Missing Values Using Neural Networks in Medical Field. 2018 *IEEE* 22*nd International Conference on Intelligent Engineering Systems* (*INES*), Las Palmas de Gran Canaria, 21-23 June 2018 151-156. https://doi.org/10.1109/ines.2018.8523857

[9] Nicholson, J.S., Deboeck, P.R. and Howard, W. (2017) Attrition in Developmental Psychology: A Review of Modern Missing Data Reporting and Practices. *International Journal of Behavioral Development*, **41**, 143-153. https://doi.org/10.1177/0165025415618275

[10] Chourib, I., Guillard, G., Mestiri, M., Solaiman, B. and Farah, I.R. (2020) Case-Based Reasoning: Problems and Importance of Similarity Measure. 2020 5*th International Conference on Advanced Technologies for Signal and Image Processing* (*ATSIP*), Sousse, 2-5 September 2020, 1-6. https://doi.org/10.1109/atsip49331.2020.9231755

[11] He, Y., Zhang, G. and Hsu, C. (2021) Multiple Imputation of Missing Data in Practice. Chapman and Hall/CRC. https://doi.org/10.1201/9780429156397

[12] Allison and Paul D (2001) Sage University Papers Series on Quantitative Applications in the Social Sciences. Sage, 136.

[13] Gelman, A. and Hill, J. (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. https://doi.org/10.1017/cbo9780511790942

[14] Batista, G.E.A.P.A. and Monard, M.C. (2003) Experimental Comparison of K-Nearest Neighbor and Mean or Mode Imputation Methods with the Internal Strategies

Used by C4.5 and CN2 to Treat Missing Data. University of Sao Paulo, 34.

[15] Zhang, Z. (2016) Missing Data Imputation: Focusing on Single Imputation. *Annals of Translational Medicine*, **4**, Article 9.

[16] Farhangfar, A., Kurgan, L. and Dy, J. (2008) Impact of Imputation of Missing Values on Classification Error for Discrete Data. *Pattern Recognition*, **41**, 3692-3705. https://doi.org/10.1016/j.patcog.2008.05.019

[17] Uenal, H., Mayer, B. and Du Prel, J.B. (2014) Choosing Appropriate Methods for Missing Data in Medical Research: A Decision Algorithm on Methods for Missing Data. *Journal of Applied Quantitative Methods*, 9.

[18] Mattei, P.-A. and Frellsen, J. (2019) MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. *Proceedings of the* 36*th International Conference on Machine Learning*, **97**, 4413-4423.

[19] Liu, Y., Cui, P., Hu, W. and Hong, R. (2023) Deep Ensembles Meets Quantile Regression: Uncertainty-Aware Imputation for Time Series. arXiv: 2312.01294.

[20] Hernandez-Lobato, Miguel, J., Zhang, Y. and Zoubin, G. (2022) Bayesian Regression Imputation for Missing Data with Uncertainty Quantification. *Journal of Machine Learning Research*, **23**, 1-30.

[21] Cover, T. and Hart, P. (1967) Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **13**, 21-27. https://doi.org/10.1109/tit.1967.1053964

[22] Duy Le, T., Beuran, R. and Tan, Y. (2018) Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. 2018 10*th International Conference on Knowledge and Systems Engineering* (*KSE*), Ho Chi Minh City, 1-3 November 2018, 247-251. https://doi.org/10.1109/kse.2018.8573344

[23] Kim, Y. and Lee, J. and Park, S. (2023) SAITS: Self-Attention Imputation for Time Series. *Neurocomputing*, **523**, 19-30.

[24] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[25] Zakkour, A., Perret, C. and Slaoui, Y. (2023) Stochastic Expectation Maximization Algorithm for Linear Mixed-Effects Model with Interactions in the Presence of Incomplete Data. *Entropy*, **25**, Article 473. https://doi.org/10.3390/e25030473

[26] Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., *et al.* (2010) Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem. *Artificial Intelligence in Medicine*, **50**, 105-115. https://doi.org/10.1016/j.artmed.2010.05.002

[27] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004) Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmospheric Environment*, **38**, 2895-2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

[28] Tang, F. and Ishwaran, H. (2017) Random Forest Missing Data Algorithms. *Statistical Analysis and Data Mining*: *The ASA Data Science Journal*, **10**, 363-377. https://doi.org/10.1002/sam.11348

[29] Aleryani, A., Wang, W. and De La Iglesia, B. (2018) Dealing with Missing Data and Uncertainty in the Context of Data Mining. In: de Cos Juez, F., *et al.*, Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 289-301. https://doi.org/10.1007/978-3-319-92639-1_24

[30] Yoon, J., Jordon, J. and Schaar, M. (2018) GAIN: Missing Data Imputation Using Generative Adversarial Nets. *Proceedings of the* 35*th International Conference on Machine Learning*, Vol. 80, 5689-5698.

[31] Cao, W., Wang, D., Li, J., Zhou, H., Li, L. and Li, Y. (2018) BRITS: Bidirectional Recurrent Imputation for Time Series. *Advances in Neural Information Processing Systems*, **31**, 6775-6785.

[32] Du, W., Côté, D. and Liu, Y. (2023) SAITS: Self-Attention-Based Imputation for Time Series. *Expert Systems with Applications*, **219**, Article ID: 119619.

[33] Xu, Y., Li, Q. and Gao, J. (2022) Miss Former: Transformer-Based Missing Value Imputation. arXiv: 2205.04296.

[34] Cini, A., Marisca, I. and Alippi, C. (2022) Filling the Gaps: Multi-Variate Time Series Imputation by Graph Neural Networks. *International Conference on Learning Representations*. arXiv: 2108.00298.

[35] Gordon, D., Petousis, P., Zheng, H., Zamanzadeh, D. and Bui, A.A.T. (2021) TSI-GNN: Extending Graph Neural Networks to Handle Missing Data in Temporal Settings. *Frontiers in Big Data*, **4**, Article 693869. https://doi.org/10.3389/fdata.2021.693869

[36] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[37] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, We., Ye, Q. and Liu, T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3146-3154.

[38] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*, **31**, 6638-6648.

[39] Chai, T. and Draxler, R.R. (2014) Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development*, **7**, 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

[40] Mesquita, D.P.P., Gomes, J.P.P., Souza Junior, A.H. and Nobre, J.S. (2017) Euclidean Distance Estimation in Incomplete Datasets. *Neurocomputing*, **248**, 11-18. https://doi.org/10.1016/j.neucom.2016.12.081

[41] McHugh, M.L. (2012) Interrater Reliability: The Kappa Statistic. *Biochemia Medica*, **22**, 276-282. https://doi.org/10.11613/bm.2012.031