Journal of Computer and Communications, 2024, 12, 53-75
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Missing Data Imputation: A Comprehensive Review

Majed Alwateer[1], El-Sayed Atlam[1,2], Mahmoud Mohammed Abd El-Raouf[3], Osama A. Ghoneim[4], Ibrahim Gad[2]

[1]Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu, Saudi Arabia
[2]Computer Science Department, Faculty of Science, Tanta University, Tanta, Egypt
[3]Basic and Applied Science Institute, College of Engineering and Technology, Arab Academy for Science and Technology (AAST), Alexandria, Egypt
[4]Department of Computer Science, Faculty of Computers and informatics, Tanta University, Tanta, Egypt
Email: satlam@taibahu.edu.sa, mwateer@taibahu.edu.sa, ibrahim.gad@science.tanta.edu.eg, m_abdelraouf85@aast.edu

How to cite this paper: Alwateer, M., Atlam, E.-S., Abd El-Raouf, M.M., Ghoneim, O.A. and Gad, I. (2024) Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications*, **12**, 53-75.
https://doi.org/10.4236/jcc.2024.1211004

Received: September 25, 2024
Accepted: November 8, 2024
Published: November 11, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Missing data presents a significant challenge in statistical analysis and machine learning, often resulting in biased outcomes and diminished efficiency. This comprehensive review investigates various imputation techniques, categorizing them into three primary approaches: deterministic methods, probabilistic models, and machine learning algorithms. Traditional techniques, including mean or mode imputation, regression imputation, and last observation carried forward, are evaluated alongside more contemporary methods such as multiple imputation, expectation-maximization, and deep learning strategies. The strengths and limitations of each approach are outlined. Key considerations for selecting appropriate methods, based on data characteristics and research objectives, are discussed. The importance of evaluating imputation's impact on subsequent analyses is emphasized. This synthesis of recent advancements and best practices provides researchers with a robust framework for effectively handling missing data, thereby improving the reliability of empirical findings across diverse disciplines.

Keywords

Missing Data, Machine Learning, Prediction, Deep Learning, Imputation

1. Introduction

Missing data can occur due to various reasons, such as participant non-response in surveys, equipment malfunction in experimental settings, or data entry errors

[1] [2]. The presence of missing data can significantly impact statistical analyses and lead to incorrect conclusions if not properly addressed [3]. Missing data is a pervasive issue in empirical research across various disciplines, including social sciences, medical research, and data science [4]. It occurs when no value is available for a variable in an observation, potentially leading to incomplete datasets that can compromise the validity and reliability of statistical analyses [5] [6].

Missing values can significantly impact your analysis: They can introduce bias if not handled properly. Many machine learning algorithms can't handle missing values that are out of the box. They can lead to the loss of important information if instances with missing values are simply discarded. Improperly handled missing values can lead to incorrect conclusions or predictions.

Missing values can sneak into your data for a variety of reasons. Here are some common reasons: Data Entry Errors: Sometimes, it's just human error. Someone might forget to input a value or accidentally delete one. Sensor Malfunctions: In IoT or scientific experiments, a faulty sensor might fail to record data at certain times [7]. Survey Non-Response: In surveys, respondents might skip questions they are uncomfortable answering or don't understand. Merged Datasets: When combining data from multiple sources, some entries might not have corresponding values in all datasets. Data Corruption: During data transfer or storage, some values might get corrupted and become unreadable. Intentional Omissions: Some data might be intentionally left out due to privacy concerns or irrelevance. Sampling Issues: The data collection method might systematically miss certain types of data. Time-Sensitive Data: In time series data, values might be missing for periods when data wasn't collected (e.g., weekends, holidays) [8].

Researchers typically encounter three main types of missing data mechanisms, as defined by Rubin [4]: 1) Missing Completely at Random (MCAR): The probability of missing data is unrelated to both observed and unobserved variables. 2) Missing at Random (MAR): The probability of missing data depends on observed variables but not on unobserved variables. 3) Missing Not at Random (MNAR): The probability of missing data depends on unobserved variables, including the missing data itself.

Addressing missing data is of paramount importance in research for several reasons: 1) Bias reduction: Ignoring missing data or using simplistic methods like complete case analysis can lead to biased estimates and incorrect inferences [9]. Proper handling of missing data helps minimize this bias and improve the accuracy of research findings. 2) Statistical power: Missing data reduces the effective sample size, leading to decreased statistical power. Appropriate imputation techniques can help maintain or even increase power by utilizing all available information [10]. 3) Generalizability: Incomplete datasets may not be representative of the population of interest, potentially limiting the generalizability of research findings. Addressing missing data can help improve the external validity of results [11]. 4) Ethical considerations: In clinical trials and other human subjects research, ignoring missing data may lead to the waste of valuable data that participants have provided, raising ethical concerns. 5) Regulatory compliance: In some fields, such as clinical trials, regulatory bodies require proper handling and reporting

of missing data [12]. 6) Improved decision-making: In applied settings, such as business analytics or public policy, accurate and complete data is essential for informed decision-making [13].

While imputation methods offer valuable tools for handling missing data, several challenges and considerations must be addressed to ensure effective and reliable results: 1) Handling different data types: Imputation methods must be able to handle various data types, including continuous, categorical, and mixed data [14]. 2) Dealing with high-dimensional data: As the number of variables increases, imputation becomes more challenging due to the curse of dimensionality [15]. 3) Computational efficiency: Some advanced imputation methods can be computationally intensive, necessitating efficient implementations for large-scale applications [16]. 4) Preserving data distributions and relationships: Imputation methods should maintain the statistical properties of the original data, including distributions and relationships between variables [17].

Missing data is a pervasive issue in research across various disciplines, often leading to biased or inefficient analyses [4]. Addressing missing data is crucial for maintaining the integrity and reliability of research findings [18]. This review aims to provide a comprehensive overview of missing data imputation techniques, their applications, and current challenges in the field.

Table 1 summarizing the recent papers on imputation methods published between 2017 and 2024, focusing on the year, model used, columns imputed, and key results.

Given the critical nature of missing data in research, this comprehensive review aims to achieve the following objectives: 1) Provide an up-to-date synthesis of current missing data imputation techniques, including traditional methods and advanced machine learning approaches. 2) Critically evaluate the strengths and limitations of various imputation methods across different research contexts and data types. 3) Examine the impact of different missing data mechanisms on the performance of imputation techniques. 4) Explore emerging trends and future directions in missing data imputation, including the application of deep learning and artificial intelligence techniques.

Offer practical guidelines to help researchers select appropriate imputation methods based on their specific research context, data characteristics, and analysis goals. Identify gaps in the current literature and propose areas for future research in missing data imputation.

- Provide an overview of traditional and advanced imputation methods.
- Discuss evaluation metrics for imputation techniques.
- Examine challenges and considerations in missing data imputation.
- Explore case studies and applications across various domains.
- Identify future directions and open problems in the field data imputation.

By addressing these objectives, this review aims to provide researchers, statisticians, and data scientists with a comprehensive understanding of missing data imputation techniques, their applications, and their implications for research integrity and validity.

Table 1. Summary of imputation methods in published papers (2017-2024).

Paper title	Year	Model used	Columns imputed	Key results
Doreswamy *et al.* [19]	2017	kernel ridge, linear regression, random forest, SVM, and KNN	Multiple variables of NCDC weather dataset	Accounted for temporal dependencies in longitudinal data, leading to more accurate estimates of parameters and improved model performance.
Hosahalli *et al.* [20]	2018	Machine learning models	NCDC weather datasets	Improved accuracy of predictive models for NCDC weather dataset compared to single imputation methods.
Khanani [21]	2021	Predictive mean matching (PMM)	Education data	Demonstrated effectiveness in imputing missing values in educational data from a public school with both numerical and categorical features.
Thakur *et al.* [22]	2021	Machine learning	Time series data	Provided a comprehensive overview of multiple imputation techniques and their applications in machine learning, highlighting their advantages and limitations.
Psychogyios *et al.* [23]	2023	KNN-MICE-GAN	Age, gender, diagnosis codes, lab results	Improved accuracy of predictive models for hospital readmission compared to single imputation methods.
Omar *et al.* [24]	2023	Random forest, decision tree, neural network and support vector machine	Dropout in higher education	The results showed that the Random Forest algorithm obtained the best performance, with an AUC of 0.9623 in the prediction of college dropout.
Nida *et al.* [25]	2023	Mean imputation, KNN, PMM	Rainfall data	The KNN achieved high imputation accuracy for missing rainfall variables values, improving the analysis of complex weather datasets.
Psychogyios *et al.* [23]	2023	GAN	Electronic health records	The results show that GAN achieved high imputation accuracy and outperform the standard baselines.
Teegavarapu *et al.* [26]	2024	Spatial and temporal interpolation methods	Hydrometeorological data	Provided a comprehensive overview of multiple imputation techniques for precipitation, temperature, and streamflows and highlighting their advantages and limitations.
Almeida *et al.* [27]	2024	Focalize K-NN method	Time series data	The results demonstrated that the effectiveness of Focalize K-NN for imputing missing values in time series data.
Kowsar *et al.* [28]	2024	Self-attention imputation method	Electronic health records	The proposed imputation method demonstrates superior performance across a range of missing data proportions (10% to 50%) under the assumption of missing completely at random (MCAR).

2. Missing Data Types

Understanding the mechanisms of missingness is essential for selecting appropriate imputation methods and accurately interpreting results. Rubin's classification system for missing data, which remains fundamental to the field, identifies three main types of missing data [18].

2.1. Missing Completely at Random (MCAR)

Missingness is considered completely random when it does not depend on any other variables. This condition, known as Missing Completely at Random (MCAR), occurs when the probability of missing data is unrelated to both observed and unobserved variables [29]-[31]. In this scenario, the missingness is purely due to chance, with no influence from any characteristics of the data [4]. MCAR is the most stringent assumption and rarely occurs in practice. However, when data are MCAR, analyses using only complete cases will be unbiased, although potentially inefficient [1].

Let R be the missingness indicator and Y be the complete data. MCAR can be defined as: $P(R \mid Y) = P(R)$.

For example, in a survey, some participants accidentally skip questions regardless of their characteristics or responses to other items.

2.2. Missing at Random (MAR)

The probability of missing data depends on other observed variables but not on the missing data itself. This condition, known as Missing at Random (MAR), occurs when the likelihood of missingness is related to observed variables but not to unobserved ones [18]. MAR represents a less stringent assumption compared to Missing Completely at Random (MCAR) and is often more realistic in practice. In summary, data are considered MAR when the probability of missingness is influenced by observed variables while remaining independent of unobserved variables [18]. MAR allows for relationships between observed variables and the probability of missingness. Many modern imputation methods, such as multiple imputation, assume MAR [17].

Let Y_{obs} be the observed data and Y_{mis} be the missing data. MAR can be defined as: $P(R \mid Y) = P(R \mid Y_{obs})$.

For example, in a longitudinal study on income, older participants are more likely to withhold information about their earnings, but this likelihood is not related to the actual income amount after accounting for age. For example, men might be less likely to answer questions about emotions in a survey.

2.3. Missing Not at Random (MNAR)

Missingness is categorized as Missing Not at Random (MNAR) when the probability of missing data is influenced by the values of the missing data itself or by unobserved variables. In this case, the missingness is directly related to the unobserved

values [4]. MNAR is the most challenging type of missing data to handle. Standard imputation methods can yield biased results, necessitating specialized techniques or sensitivity analyses to address the issue effectively [32].

MNAR occurs when the probability of missing data depends on unobserved variables or the missing values themselves [4]. This is the most challenging type of missingness to address and often requires specialized techniques. MNAR occurs when: $P(R|Y) \neq P(R|Y_{obs})$.

An example of MNAR is in a mental health survey, participants with severe depression may be less likely to complete questions about their symptoms, with this likelihood directly related to the severity of their undisclosed condition. Similarly, individuals with high incomes might be less inclined to report their income in a survey.

In practice, it is often impossible to definitively determine whether data are MAR or MNAR based solely on observed data [33]. Therefore, researchers often must rely on subject-matter knowledge and conduct sensitivity analyses to assess the robustness of their findings under different missing data assumptions [17].

3. Missing Value Imputation

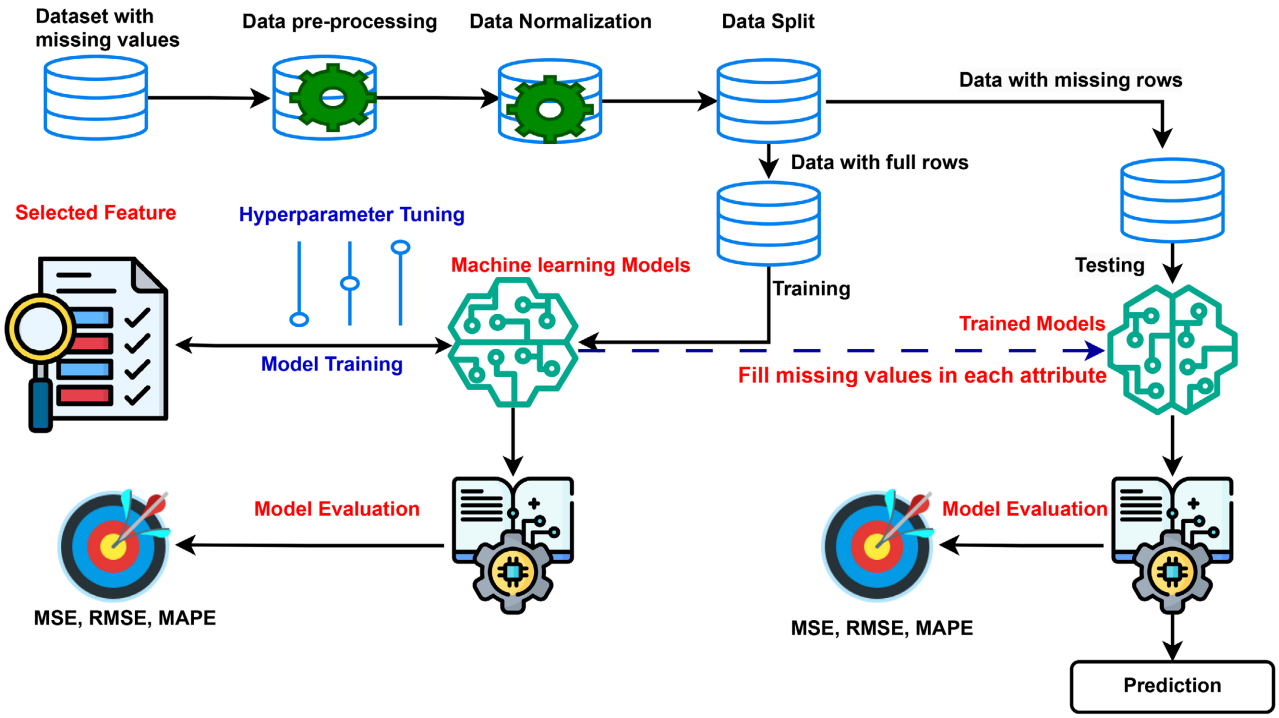

Figure 1. The flowchart of imputation of missing values.

The problem of missing data imputation arises when a dataset contains unobserved values for certain features. Let $X = (X_1, X_2, \cdots, X_n) \in \mathbb{R}^n$ represent a random vector, where each X_i corresponds to a feature and follows a distribution $P(X)$. A binary mask vector $M = (M_1, M_2, \cdots, M_n)$ indicates the presence or

absence of observations, with $M_i = 1$ denoting an observed value and $M_i = 0$ a missing value. Given a dataset of d instances $\{(X_i, M_i)\}_{i=0,\cdots,d-1}$, an imputed dataset is constructed by replacing missing values (where $M_{ij} = 0$) in the observed data X_i with pre-imputed values, potentially random noise, resulting in $\hat{X}_i$. The objective is to develop an imputation model, IMP, that generates an imputed dataset $\left\{\tilde{X}_i = IMP\left(\hat{X}_i, M_i\right)\right\}_{i=0,\cdots,d-1}$ such that each imputed sample $\tilde{X}_i$ is drawn from the conditional distribution $P(X \mid \tilde{X}_i)$, thereby preserving the original data's distributional properties. This yields a complete dataset $\bar{X}$, where $\bar{X}_i = X_i \odot M_i + (1 - M_i) \odot \tilde{X}_i$ for each sample i.

Following imputation of missing values, predictive modeling was performed as shown in Figure 1. Application of S distinct imputation algorithms, $A_0, A_1, \cdots, A_{S-1}$, to the original dataset X yielded S completed datasets, denoted $\bar{D}_1, \bar{D}_2, \cdots, \bar{D}_S$. A standard predictive model, P, was then applied to each of these completed datasets to predict the outcome.

4. Methods

This study evaluated several imputation methods for handling missing data, ranging from simple statistical techniques to more sophisticated deep learning approaches. Specific methods included.

4.1. Traditional Imputation Methods

Traditional methods for handling missing data have been widely used due to their simplicity and ease of implementation [1]. Traditional imputation methods are widely used to handle missing data in datasets. These methods aim to replace missing values with plausible estimates based on the available data. The most common traditional imputation techniques are discussed as follows.

4.1.1. Mean/Median Imputation

Missing categorical values were replaced with the mode, while missing numerical values were replaced with the mean of the corresponding feature. This method involves replacing missing values with the mean or median of the observed values for that variable [4]. Although simple, this approach can result in biased estimates and an underestimation of standard errors. For numerical data, missing values are replaced with the mean or median of the non-missing values in the same column, while for categorical data, the mode (most frequent value) is utilized.

4.1.2. MissForest

This iterative approach uses mean/mode imputation to initialize the dataset [16]. Then, a random forest is trained to predict the missing values in each feature, iteratively refining the imputed values until convergence (defined by a lack of improvement in the imputed matrix). Convergence criteria included a maximum of 20 iterations and 100 trees. The difference between successive imputed matrices (M_{imp_new} and M_{imp_old}) for numerical (N) and categorical (F) features was measured as:

$$\delta_N = \frac{\sum_{j \in N} \left(M_{new}^{imp} - M_{old}^{imp} \right)^2}{\sum_{j \in N} \left(M_{new}^{imp} \right)^2} \quad (1)$$

$$\delta_F = \frac{\sum_{j \in F}^{} \sum_{i=1}^{i=n} V_{M_{new}^{imp} \neq M_{old}^{imp}}}{F_{NA}} \quad (2)$$

where F_{NA} represents the number of missing values in categorical variables.

4.1.3. LOCF and NOCB

Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB) are both methods used for handling missing data in longitudinal studies. LOCF fills in missing data points by carrying forward the last observed value. For example, if a participant's value is missing at a follow-up, the last recorded value is used to fill in that gap [34].

The advantages of LOCF method are: 1) Simplicity: Easy to implement and understand. 2) Preservation of Sample Size: Retains all participants in the analysis, which can be important in clinical trials. On the other hand, the disadvantages of LOCF method are: 1) Assumption of Stability: Implies that the last observation is a good estimate for future values, which may not hold true. 2) Potential Bias: Can introduce bias if the last observed value is not representative of the participant's state at the time of the missing data. 3) Underestimation of Variability: Fails to account for natural fluctuations in the data, potentially leading to misleading conclusions.

NOCB fills in missing data points by carrying the next observed value backward. For example, if a participant's value is missing before a subsequent observation, the next recorded value is used to fill in the gap.

The advantages of NOCB method are: 1) Preservation of Trends: Can better reflect changes over time if later observations are more representative of the participant's condition. 2) Potentially Reduces Bias: Addresses some issues associated with LOCF by using future data, which may be more accurate. On the other hand, the disadvantages of NOCB method are: 1) Assumption of Continuity: Assumes that the value observed in the future can be reliably applied to the past, which may not always be valid. 2) Temporal Distortion: Can introduce bias if there are systematic changes between the missing data point and the next observation. 3) More Complex: Generally considered less intuitive and harder to justify in some contexts than LOCF.

4.1.4. Hot Deck Imputation

Hot deck imputation involves replacing missing values with observed values from similar respondents or cases [35]. This method can help preserve the distribution of the data but may be challenging to implement for large datasets. This method replaces missing values with values from a similar donor record (a record with non-missing values) in the dataset.

The main steps of Hot Deck Imputation are: 1) Define a set of matching criteria

(e.g., age, gender, income) based on the variables with available data. 2) For each missing value, find a donor record that matches the criteria. 3) Replace the missing value with the corresponding value from the donor record. Advantages: Simple, can be effective for handling missing values in categorical variables.

4.1.5. Multivariate Imputation by Chained Equations (MICE)

This method generates n imputed datasets [36]. Parameter estimates and standard errors are calculated for each dataset, and then pooled to obtain overall estimates ($\bar{P}$) and variances ($\bar{V}$):

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$$

Between-dataset variability ($\bar{B}$) is also calculated:

$$\bar{B} = \frac{1}{n-1}\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2$$

4.1.6. Neighborhood Aware Autoencoder (NAA)

This approach uses a denoising autoencoder, pre-imputed with kNN ($k = 5$), to learn feature relationships and impute missing values [37]. The encoder and decoder are defined by:

$$f_{enc}\left(\mathbf{X}\right) = s\left(\mathbf{X}\cdot W^{\mathrm{T}} + b\right)$$

$$f_{dec}\left(\mathbf{Y}\right) = s\left(\mathbf{Y}\cdot \bar{W}^{\mathrm{T}} + \bar{b}\right)$$

where $\mathbf{Y} = f_{enc}\left(\mathbf{X}\right)$ and $\mathbf{Z} = f_{dec}\left(\mathbf{Y}\right)$ are the hidden and output vectors, respectively; W and b are the encoder weights and bias; and $\bar{W}$ and $\bar{b}$ are the decoder weights and bias. Training minimizes the reconstruction error between $\mathbf{X}$ and $\mathbf{Z}$.

4.1.7. Improved Neighborhood Aware Autoencoder (I-NAA)

This enhanced version uses an undercomplete autoencoder architecture. To avoid overfitting to the initial kNN imputation, the kNN imputation is updated every 10 epochs, varying the k value within a predefined range. Furthermore, the missing values to be imputed are randomly selected at the start of each epoch. A custom loss function combines mean squared error (MSE) for numerical features and binary cross-entropy (BCE) for categorical features:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2$$

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=N+1}^{N+C} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1-y_i\right)\cdot\log\left(1-p\left(y_i\right)\right)$$

$$\mathrm{Loss} = \mathrm{RMSE} + \mathrm{BCE}$$

4.1.8. Multiple Imputation (MI)

Multiple imputation (MI) is a powerful technique for handling missing data that addresses the limitations of single imputation methods. Unlike single imputation, which replaces missing values with a single estimate, MI generates multiple complete datasets by imputing missing values multiple times, each time using different plausible values [38]. This approach accounts for the uncertainty introduced by missing data, leading to more accurate and robust analyses.

The main steps of MI method are: 1) Imputation: Multiple complete datasets are created by imputing the missing values using a statistical model that accounts for the relationships between variables. The model is typically based on the observed data and assumes a specific distribution for the missing values. Each imputed dataset is generated using different random draws from the conditional distribution of the missing values, reflecting the uncertainty associated with the missing data. 2) Analysis: Each of the imputed datasets is analyzed separately using the chosen statistical methods. This results in multiple sets of estimates for the parameters of interest. 3) Pooling: The results from each imputed dataset are combined using appropriate methods to obtain a single set of estimates and standard errors that reflect the uncertainty introduced by missing data. The most common pooling methods include averaging the estimates and variances across the imputed datasets.

The advantages of Multiple Imputation are: 1) Accounts for Uncertainty: MI explicitly acknowledges the uncertainty associated with missing values by generating multiple plausible estimates. This results in more realistic confidence intervals and p-values [39]. 2) Reduces Bias: By generating multiple imputed datasets, MI reduces the bias introduced by single imputation methods, especially when the missing data is not missing at random. 3) More Accurate Estimates: MI generally produces more accurate estimates of parameters and statistical tests than single imputation methods. 4) Provides Insights into Missing Data: The variability of estimates across imputed datasets can provide insights into the sensitivity of the analysis to the missing data.

The challenges of Multiple Imputation are: 1) Computational Complexity: MI can be computationally intensive, especially for large datasets and complex models. 2) Model Selection: Choosing the appropriate imputation model is crucial. The model should accurately reflect the relationships between variables and the distribution of the missing data. 3) Software Requirements: Specialized software is often required to perform multiple imputation, as it involves generating and analyzing multiple datasets.

The choice of imputation method depends on the specific characteristics of the data, the nature of the missing data, and the goals of the analysis. It is important to consider the potential biases and limitations of each method before applying it to your data. **Table 2** summarizing common imputation methods, their use cases, advantages, disadvantages, Python packages, and suitability for classification or regression problems.

Table 2. Summary of imputation methods.

Method	Use cases	Advantages	Disadvantages	Python package	Problem type
Mean/Median Imputation	Simple missing value replacement, suitable for numerical data with a clear central tendency.	Simple, computationally inexpensive.	Can introduce bias, especially for non-normally distributed data. Does not account for relationships between variables.	"SimpleImputer" (scikit-learn)	Regression
K-Nearest Neighbors (KNN)	Handles both numerical and categorical data, accounts for relationships between variables.	Accounts for relationships between variables, effective for both numerical and categorical data.	Can be computationally expensive for large datasets, sensitive to the choice of k.	"KNNImputer" (scikit-learn)	Regression/ Classification
Last Observation Carried Forward (LOCF)	Primarily for time series data, replaces missing values with the last observed value.	Simple, can be effective for time series data with a strong trend.	Can introduce bias if the data is not trending, can propagate errors if there are consecutive missing values.	"fillna (method = 'ffill')" (pandas)	Time Series
Multiple Imputation (MI)	Handles complex missing data patterns, accounts for uncertainty in imputation.	Accounts for uncertainty in imputation, can provide more accurate estimates than single imputation methods.	Can be computationally expensive, requires specialized software.	"IterativeImputer" (scikit-learn), "fancyimpute"	Regression/ Classification
Hot-Deck Imputation	Primarily for categorical data, replaces missing values with values from a similar donor record.	Simple, can be effective for handling missing values in categorical variables.	Can introduce bias if the donor records are not truly similar to the record with the missing value.	"KNNImputer" (scikit-learn) can be adapted	Classification

4.2. Advanced Imputation Techniques

Regression imputation uses the relationship between variables to predict missing values based on observed data [9]. This method can account for relationships between variables but may overestimate the strength of these relationships.

4.2.1. K-Nearest Neighbors (KNN) Imputation

This method identifies the k-nearest neighbors to the missing value based on the similarity of other features [40]. The missing value is then replaced with the average (for numerical data) or the most frequent value (for categorical data) among those neighbors.

The main steps of KNN Imputation are: 1) Calculate the distance between the data point with the missing value and all other points in the dataset. 2) Identify the k-nearest neighbors based on these distances. 3) For numerical data, compute the average of the corresponding values in the k-nearest neighbors and use it as the imputed value. For categorical data, select the most frequent value among the

neighbors.

Missing values were imputed using the average (numerical features) or mode (categorical features) of the k nearest neighbors in feature space, using Euclidean distance. For example, $k=4$. Formally, for a sample $S(X,Y,0)$ with four nearest neighbors $N_4=\{(X_i,Y_i,1)\,|\,i=1,2,3,4\}$, the imputed value Y is calculated as:

$$Y=\begin{cases}\arg\max_z\left\{\sum_{(X_i,Y_i,1)\in N_4}1(Y_i=z)\right\} & \text{if } Y \text{ is categorical}\\ \frac{1}{4}\sum_{i=1}^{4}Y_i & \text{if } Y \text{ is numerical}\end{cases}$$

where $z\in\{0,1\}$ and $1(Y_i=z)$ is an indicator function. The Euclidean distance between points x and y is defined as:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

where n is the number of features.

KNN imputation identifies the k most similar cases to the ones with missing data and uses their values for imputation [41] [42]. This method can capture complex relationships in the data but may be computationally expensive for large datasets. K-Nearest Neighbors (KNN) imputation is a popular method for handling missing data, leveraging the similarities between observations. Here is a closer look at how KNN imputation works, its advantages, and its limitations.

The advantages of KNN Imputation are: 1) Flexibility: KNN can be applied to both numerical and categorical data, making it versatile. 2) Local Information: By considering the closest observations, KNN can capture local data patterns, potentially leading to more accurate imputations. 3) Non-parametric: KNN does not assume a specific data distribution, which can be advantageous in real-world datasets.

The limitations of KNN Imputation are: 1) Computationally Intensive: KNN can be slow, especially with large datasets, since it requires distance calculations for each observation. 2) Curse of Dimensionality: As the number of features increases, the concept of "closeness" can become less meaningful, making it harder to identify true neighbors. Sensitive to Outliers: The presence of outliers can skew distance calculations, leading to poor imputation results.

4.2.2. Decision Trees and Random Forests

These methods use tree-based models to predict missing values based on other variables [16]. They can handle both categorical and continuous variables and capture non-linear relationships. Decision Trees and Random Forests are powerful machine learning techniques that can also be used to input missing values into datasets. Here's a breakdown of how these methods work for imputation, along with their advantages and disadvantages.

The advantages of Decision Trees Imputation: 1) Captures Non-linear Relationships: Decision trees can model complex relationships between features, potentially leading to more accurate imputations. 2) Interpretable: The model is relatively

easy to interpret, as you can visualize how decisions are made. The disadvantages of Decision Trees Imputation: 1) Overfitting: Decision trees can easily overfit to the training data, especially if not properly pruned. 2) Sensitivity to Noise: Outliers can affect the structure of the tree, impacting the imputation results.

The advantages of Random Forests Imputation: 1) Improved Accuracy: Random forests generally provide better accuracy than single decision trees due to their ensemble nature, reducing overfitting and variance. 2) Robust to Outliers: The averaging mechanism makes random forests less sensitive to outliers compared to individual trees. 3) Handles Large Datasets: Random forests can effectively manage large datasets with high dimensionality. The disadvantages of Random Forests Imputation: 1) Complexity: The model is less interpretable than a single decision tree, as it's harder to visualize how predictions are made. 2) Computationally Intensive: Training multiple trees can be resource-intensive, especially for large datasets.

4.2.3. Support Vector Machines (SVM)

Support Vector Machines (SVM) are primarily known for classification and regression tasks. However, they can also be utilized to input missing data. SVM-based imputation methods use support vector regression to predict missing values [43] [44]. These methods can be effective for high-dimensional data but may require careful tuning of hyperparameters.

The advantages of SVM-Based Imputation are: 1) Effective for Non-linear Relationships: SVM can capture complex, non-linear relationships in the data by using different kernel functions (e.g., polynomial, radial basis function). 2) Robustness to Overfitting: SVM includes regularization parameters that help prevent overfitting, making it suitable for high-dimensional datasets. 3) Flexibility: SVM can be applied to both classification (categorical variables) and regression (continuous variables) tasks.

The limitations of SVM-Based Imputation are: 1) Computational Complexity: SVM can be computationally intensive, particularly for large datasets, due to the optimization required for finding the best hyperplane. 2) Parameter Sensitivity: The performance of SVM can be sensitive to the choice of kernel and hyperparameters (e.g., C and gamma), requiring careful tuning. 3) Requires Sufficient Data: SVM models generally require a substantial amount of complete data to build an accurate model, which may not always be available.

4.3. Deep Learning Approaches

Autoencoders are neural networks that can learn compressed representations of data and have been applied to missing data imputation [45]-[47]. They can capture complex patterns in the data but may require large amounts of training data.

Generative Adversarial Networks (GANs)

GANs have been adapted for missing data imputation by learning to generate realistic imputed values [48] [49]. This approach can produce high-quality imputa-

tions but may be challenging to train and tune. The Generative Adversarial Imputation Network (GAIN) [50] employs a generative adversarial network (GAN) architecture. Unlike standard GANs, the discriminator in GAIN does not classify the entire generated output as real or fake; instead, it classifies each individual variable as either imputed or observed. Convergence is achieved when the generator produces imputations indistinguishable from the true data distribution. A "hint" mechanism augments the discriminator's input with partial information about the missing values (M), represented by the hint vector H. This hint is typically a proportion of M (e.g., 90% identical). The generator then learns to impute the remaining values. The original work demonstrates that insufficient hints lead to multiple optimal generator outputs.

Formally, given a random vector Z, the generator produces an imputed dataset $\tilde{X}_i$, from which $\bar{X}_i$ is derived (Equation (1)). The discriminator loss function, L_D, is defined as:

$$L_D\left(M,\bar{M},H\right)=\sum_{i:H_i=0}\left[M_i\cdot\log\left(\bar{M}_i\right)+\left(1-M_i\right)\cdot\log\left(1-M_i\right)\right]$$

where M is the true mask vector, $\bar{M}$ is the generated mask vector, and H is the hint vector. The summation is restricted to indices where $H_i=0$ to prevent overfitting to the hint. The discriminator is trained to minimize this loss:

$$\min_D-\sum_{i=1}^{batchsize}L_D\left(M_i,\bar{M}_i,H_i\right)$$

The generator loss function comprises two terms: L_G, which measures the generator's ability to deceive the discriminator, and L_M, which quantifies the accuracy of the imputation for observed values:

$$L_G\left(M,\bar{M},H\right)=-\sum_{i:H_i=0}\left(1-M_i\right)\cdot\log\left(M_i\right)$$

$$L_M\left(\tilde{X},\bar{X}\right)=-\sum_{i=1}^{d}M_i\cdot Diff\left(\tilde{X}_i,\bar{X}_i\right)$$

where d is the data dimensionality and $Diff$ is defined as:

$$Diff\left(\tilde{X},\bar{X}\right)=\begin{cases}\left(\tilde{X}_i-\bar{X}_i\right)^2 & \text{if } X_i \text{ is numerical}\\ -X_i\log\left(\bar{X}_i\right) & \text{if } X_i \text{ is binary}\end{cases}$$

The generator is trained to minimize the combined loss:

$$\min_G\sum_{i=1}^{batchsize}L_G\left(M_i,\bar{M}_i,H_i\right)+\alpha L_{M_i}\left(\tilde{X}_i,\bar{X}_i\right)$$

where α is a scaling parameter.

4.4. Time Series-Specific Methods

ARIMA Models

Autoregressive Integrated Moving Average (ARIMA) models are a class of statistical models used for analyzing and forecasting time series data. ARIMA models

can be employed to impute missing values in time series data by leveraging temporal dependencies [51] [52]. They are particularly useful when the data exhibits trends and seasonality. An ARIMA model is denoted as ARIMA (p, d, q), where "p" represents the order of the autoregressive (AR) component, "d" represents the degree of difference required to make the time series stationary, and "q" represents the order of the moving average (MA) component. The AR component models the relationship between the current observation and previous observations; the MA component models the relationship between the current observation and past forecast errors, and differencing (d) removes trends and makes the series stationary. A general ARIMA (p, d, q) model can be represented by the following equation:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t$$

where y_t is the time series at time t, B is the backshift operator ($By_t = y_{t-1}$), $\phi(B)$ is the autoregressive polynomial of order p, $\theta(B)$ is the moving average polynomial of order q, and ϵ_t is white noise. The choice of p, d, and q values is crucial for model fitting and depends on the characteristics of the specific time series being analyzed. Techniques like autocorrelation and partial autocorrelation functions (ACF and PACF) are often used to identify suitable model orders. Table 3 shows the summary of advanced imputation techniques.

Table 3. Summary of advanced imputation techniques.

Method	Advantages	Disadvantages	Python package	Problem type
Multiple Imputation by Chained Equations (MICE)	Handles complex relationships between variables, accounts for uncertainty in imputation.	Can be computationally expensive, requires careful model selection.	"IterativeImputer" (scikit-learn), "fancyimpute"	Regression/ Classification
Random Forest Imputation	Handles mixed-type data (numerical and categorical), robust to outliers.	Can be computationally expensive, may overfit if the data is highly correlated.	"MissForest"	Regression/ Classification
Generative Adversarial Networks (GANs)	Can generate realistic synthetic data, handles complex data distributions.	Requires significant computational resources, can be challenging to train.	"Tensorflow", "Pytorch"	Regression/ Classification
Deep Learning Imputation	Can capture complex non-linear relationships in the data, handles high-dimensional datasets.	Requires large amounts of data, can be computationally expensive, may overfit.	"Tensorflow", "Pytorch"	Regression/ Classification
Bayesian Imputation	Accounts for prior knowledge and uncertainty, provides probabilistic estimates.	Can be computationally intensive, requires careful model specification.	"PyMC3", "PyStan"	Regression/ Classification

5. Evaluation Metrics for Imputation Methods

Assessing the performance of imputation methods is crucial for selecting appropriate techniques [53]. Evaluating the performance of imputation methods is

crucial to ensure that the imputed data maintains the integrity and reliability of the original data. Several metrics are commonly used to assess the effectiveness of imputation techniques. The choice of evaluation metric depends on the specific objective of the imputation and the nature of the data. These evaluation metrics can be broadly categorized into two groups, as shown in Table 4.

Table 4. Regression and classification metrics.

Regression metrics		Classification metrics	
Mean Squared Error (MSE)	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Recall (Sensitivity)	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
Mean Absolute Percentage Error (MAPE)	$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\%$	F1-Score	$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

6. Challenges and Considerations

While imputation methods offer valuable tools for handling missing data, several challenges and considerations must be addressed when implementing missing data imputation to ensure effective and reliable results [54].

Bias Introduction: Imputation methods can introduce bias into the data, particularly if the missing values are not missing at random (MAR). This means that the missingness is related to the value of the missing variable itself or other variables in the dataset. Example: If missing values in income are more likely to occur for individuals with lower incomes, simply replacing them with the mean income will underestimate the true average income.

Data Distribution: Imputation methods often assume that the data follows a specific distribution (e.g., normal distribution). If the data deviates significantly from this assumption, the imputed values may not be representative. Example: Using mean imputation on a skewed distribution will result in imputed values that are biased towards the tail of the distribution.

Missing Value Patterns: The pattern of missing values can significantly impact the effectiveness of imputation methods. If missing values are clustered or follow a specific pattern, simple methods like mean imputation may not be appropriate. Example: If consecutive values are missing in a time series, LOCF or NOCB may introduce significant bias.

Computational Complexity: Some imputation methods, like multiple imputation or KNN imputation, can be computationally expensive, especially for large datasets. 1) Domain Knowledge: Incorporating domain knowledge into the imputation process can significantly improve the accuracy and relevance of the imputed

values. Example: In medical data, understanding the relationships between different variables and the potential causes of missing values can guide the choice of imputation method.

2) Model Selection: Choosing the appropriate imputation model is crucial. The choice should be based on the characteristics of the data, the nature of the missing values, and the goals of the analysis. 3) Interpretability: The interpretability of the imputed values is important for understanding the results of the analysis.

7. Case Studies and Applications

This section illustrates the practical implementation of missing data imputation techniques in a range of different fields. Examples of these applications can be found as follows.

- Healthcare: Examining approaches for resolving incomplete entries in electronic health records, with a focus on maintaining data integrity and improving diagnostic accuracy [55]-[57].
- Finance: Analysis of strategies for managing incomplete datasets in financial forecasting, specifically focusing on stock market predictions and their implications for investment decision-making [58] [59].
- Social Sciences: Investigation of techniques to mitigate the impact of non-response in survey data, exploring methods to preserve statistical validity and minimize bias in population-level inferences [60] [61].

8. Future Directions and Open Problems

Several areas for future research and development in missing data imputation include [62] [63].

1) Emerging techniques and research areas: Federated learning for privacy-preserving imputation [64] [65]. Reinforcement learning for adaptive imputation strategies [66]. Transfer learning for imputation in low-resource settings [67] [68].

2) Integration with big data and real-time systems: Developing scalable and efficient imputation methods for streaming data and large-scale datasets [69]-[71] as follows: Distributed Algorithms: Use scalable imputation algorithms that can handle large datasets efficiently. Techniques like mini-batch processing or parallel computing can be useful. Big Data Frameworks: Leverage tools like Apache Spark or Hadoop, which can process large volumes of data quickly and support machine learning libraries for imputation.

3) Ethical considerations in data imputation: Addressing potential biases and fairness issues in imputation methods, especially in sensitive applications like healthcare and criminal justice [72]-[74]. Here are some key points for these Ethics: a) Transparency: Researchers should clearly communicate how missing data will be handled, including the imputation methods used. This transparency builds trust and allows for reproducibility. b) Bias and Misrepresentation: Imputation can introduce bias if not done carefully. Researchers must consider whether the imputed data accurately reflects the underlying population or if it skews results.

c) Informed Consent: Participants should be informed about how their data, including any imputed values, will be used in research. This includes potential implications for privacy and the integrity of their responses. d) Appropriateness of Methods: Different imputation methods (mean, median, predictive modeling, etc.) have different assumptions. Choosing the right method is crucial to avoid distorting the data and the conclusions drawn from it. e) Impact on Decision-Making: The results derived from imputed data can influence policy or clinical decisions. Researchers should ensure that their imputation practices do not lead to harmful outcomes. f) Equity: Consider whether the imputation methods used could disproportionately affect certain groups. Ensuring that imputation methods do not reinforce existing inequalities is vital. g) Ethical Oversight: It's beneficial to have an ethical review process in place to assess the imputation strategies and their potential implications for participants and broader societal contexts. h) Data Integrity: Strive to maintain the integrity of the original dataset. Imputation should not compromise the authenticity of the data, and researchers should be mindful of the limitations that come with imputed values. i) Training and Expertise: Ensure that those involved in the imputation process have the necessary training and understanding of the ethical implications of their work.

9. Conclusion

As data collection and analysis continue to grow in importance across various domains, the field of missing data imputation is likely to see further advancements and innovations. This review has provided a comprehensive overview of missing data imputation techniques, from traditional methods, and statistical methods to advanced machine learning approaches. Key observations highlight: 1) The critical role of understanding missing data mechanisms. 2) The trade-offs between simple and complex imputation methods. 3) There is a need for careful evaluation and selection of imputation techniques. and 4) The potential of machine learning and deep learning approaches for handling complex missing data patterns. Future research should focus on developing more robust, efficient, and adaptable imputation methods that can handle the increasing complexity and scale of modern datasets while addressing ethical concerns and preserving data integrity.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Enders, C.K. (2022) Applied Missing Data Analysis. Guilford Publications.

[2] Mitchel, J.T., Kim, Y.J., Choi, J., Park, G., Cappi, S., Horn, D., *et al.* (2011) Evaluation of Data Entry Errors and Data Changes to an Electronic Data Capture Clinical Trial Database. *Drug Information Journal*, **45**, 421-430. https://doi.org/10.1177/009286151104500404

[3] Schafer, J.L. and Graham, J.W. (2002) Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-177. https://doi.org/10.1037//1082-989x.7.2.147

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

[5] Lazar, N.A. (2003) Statistical Analysis with Missing Data. *Technometrics*, **45**, 364-365. https://doi.org/10.1198/tech.2003.s167

[6] Hajjar, S. (2018) Statistical Analysis: Internal-Consistency Reliability and Construct Validity. *International Journal of Quantitative and Qualitative Research Methods*, **6**, 27-38.

[7] Noor, T.H., Atlam, E., Almars, A.M., Noor, A. and Malki, A.S. (2023) An IoT-Based Energy Conservation Smart Classroom System. *Intelligent Automation & Soft Computing*, **35**, 3785-3799. https://doi.org/10.32604/iasc.2023.032250

[8] Zhan, Y., Xia, Y., Liu, Y., Li, F. and Wang, Y. (2017) Incentive-Aware Time-Sensitive Data Collection in Mobile Opportunistic Crowdsensing. *IEEE Transactions on Vehicular Technology*, **66**, 7849-7861. https://doi.org/10.1109/tvt.2017.2692755

[9] Molenberghs, G. and Verbeke, G. (2013) Missing Data. In: Scott, M.A., Simonoff, J.S. and Marx, B.D., Eds., *The SAGE Handbook of Multilevel Modeling*, SAGE Publications Ltd, 403-424. https://doi.org/10.4135/9781446247600.n23

[10] Graham, J.W. (2009) Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, **60**, 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

[11] Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., *et al.* (2009) Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, **338**, b2393-b2393. https://doi.org/10.1136/bmj.b2393

[12] Masconi, K.L., Matsha, T.E., Echouffo-Tcheugui, J.B., Erasmus, R.T. and Kengne, A.P. (2015) Reporting and Handling of Missing Data in Predictive Research for Prevalent Undiagnosed Type 2 Diabetes Mellitus: A Systematic Review. *EPMA Journal*, **6**, Article No. 7. https://doi.org/10.1186/s13167-015-0028-0

[13] Ferris, J.A. (2009) Missing Data: A Gentle Introduction. *Drug and Alcohol Review*, **28**, 90-91. https://doi.org/10.1111/j.1465-3362.2008.00013_4.x

[14] Enders, C.K. (2017) Multiple Imputation as a Flexible Tool for Missing Data Handling in Clinical Research. *Behaviour Research and Therapy*, **98**, 4-18. https://doi.org/10.1016/j.brat.2016.11.008

[15] Bertsimas, D., Pawlowski, C. and Zhuo, Y.D. (2018) From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*, **18**, 1-39.

[16] Stekhoven, D.J. and Bühlmann, P. (2011) Missforest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, **28**, 112-118. https://doi.org/10.1093/bioinformatics/btr597

[17] Van Buuren, S. (2018) Flexible Imputation of Missing Data. CRC Press.

[18] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.2307/2335739

[19] Doreswamy, Gad, I. and Manjunatha, B.R. (2017) Performance Evaluation of Predictive Models for Missing Data Imputation in Weather Data. 2017 *International Conference on Advances in Computing, Communications and Informatics* (*ICACCI*), Udupi, 13-16 September 2017, 1327-1334. https://doi.org/10.1109/icacci.2017.8126025

[20] Hosahalli, D. and Gad, I. (2018) A Generic Approach of Filling Missing Values in NCDC Weather Stations Data. 2018 *International Conference on Advances in*

Computing, Communications and Informatics (*ICACCI*), Bangalore, 19-22 September 2018, 143-149. https://doi.org/10.1109/icacci.2018.8554394

[21] Khanani, N. (2021) Addressing Missing Data in Educational Evaluation: Predictive Mean Matching Imputation for Test Score Data. *Proceedings of the* 2021 *AERA Annual Meeting*, 4 September 2021, 15. https://doi.org/10.3102/1687298

[22] Thakur, S., Choudhary, J. and Singh, D.P. (2021) A Survey on Missing Values Handling Methods for Time Series Data. In: Sheth, A., Sinhal, A., Shrivastava, A. and Pandey, A.K., Eds., *Intelligent Systems*, Springer Singapore, 435-443. https://doi.org/10.1007/978-981-16-2248-9_42

[23] Psychogyios, K., Ilias, L., Ntanos, C. and Askounis, D. (2023) Missing Value Imputation Methods for Electronic Health Records. *IEEE Access*, **11**, 21562-21574. https://doi.org/10.1109/access.2023.3251919

[24] Jiménez, O., Jesús, A. and Wong, L. (2023) Model for the Prediction of Dropout in Higher Education in Peru Applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine. 2023 33*rd Conference of Open Innovations Association* (*FRUCT*), Zilina, 24-26 May 2023, 116-124. https://doi.org/10.23919/fruct58615.2023.10143068

[25] Nida, H. (2023) Comparison of Missing Data Imputation Methods Using Weather Data. *Pakistan Journal of Agricultural Sciences*, **60**, 327-336. https://doi.org/10.21162/pakjas/23.228

[26] Teegavarapu, R.S.V. (2024) Applications: Imputation of Missing Hydrometeorological Data. *Water Science and Technology Library*, **108**, 491-517. https://doi.org/10.1007/978-3-031-60946-6_8

[27] Almeida, A., Brás, S., Sargento, S. and Pinto, F.C. (2024) Focalize K-NN: An Imputation Algorithm for Time Series Datasets. *Pattern Analysis and Applications*, **27**, Article No. 39. https://doi.org/10.1007/s10044-024-01262-3

[28] Kowsar, I., Rabbani, S.B. and Samad, M.D. (2024) Attention-Based Imputation of Missing Values in Electronic Health Records Tabular Data. 2024 *IEEE* 12*th International Conference on Healthcare Informatics* (*ICHI*), Orlando, 3-6 June 2024, 177-182. https://doi.org/10.1109/ichi61247.2024.00030

[29] Little, R.J.A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, **83**, 1198-1202. https://doi.org/10.2307/2290157

[30] Hashim, H., Almaliki, M., El-Agamy, R., El Sharkasy, M., Dagnew, G., Gad, I., Ghoneim, O., *et al.* (2021) Integrating Data Warehouse and Machine Learning to Predict on COVID-19 Pandemic Empirical Data. *Journal of Theoretical and Applied Information Technology*, **99**, 159-170.

[31] Malki, Z., Atlam, E., Hassanien, A.E., Dagnew, G., Elhosseini, M.A. and Gad, I. (2020) Association between Weather Data and COVID-19 Pandemic Predicting Mortality Rate: Machine Learning Approaches. *Chaos, Solitons & Fractals*, **138**, Article ID: 110137. https://doi.org/10.1016/j.chaos.2020.110137

[32] Laaksonen, S. (2018) Survey Methodology and Missing Data: Tools and Techniques for Practitioners. Springer. https://doi.org/10.1007/978-3-319-79011-4

[33] Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, 3-15. https://doi.org/10.1191/096228099671525676

[34] Lee, Y. (2023) Imputation Method Using Local Linear Regression Based on Bidirectional *k*-nearest-components. *Journal of information and communication convergence engineering*, **21**, 62-67. https://doi.org/10.56977/jicce.2023.21.1.62

[35] Andridge, R.R. and Little, R.J.A. (2010) A Review of Hot Deck Imputation for Survey Non-Response. *International Statistical Review*, **78**, 40-64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

[36] van Buuren, S. (2007) Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, **16**, 219-242. https://doi.org/10.1177/0962280206074463

[37] Aidos, H. and Tomas, P. (2021) Neighborhood-Aware Autoencoder for Missing Value Imputation. 2020 28*th European Signal Processing Conference* (*EUSIPCO*), Amsterdam, 18-21 January 2021, 1542-1546. https://doi.org/10.23919/eusipco47968.2020.9287580

[38] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. Wiley. https://doi.org/10.1002/9780470316696.ch5

[39] Rombach, I., Gray, A.M., Jenkinson, C., Murray, D.W. and Rivero-Arias, O. (2018) Multiple Imputation for Patient Reported Outcome Measures in Randomised Controlled Trials: Advantages and Disadvantages of Imputing at the Item, Subscale or Composite Score Level. *BMC Medical Research Methodology*, **18**, Article No. 87. https://doi.org/10.1186/s12874-018-0542-6

[40] Murti, D.M.P., Pujianto, U., Wibawa, A.P. and Akbar, M.I. (2019) K-Nearest Neighbor (K-NN) Based Missing Data Imputation. 2019 5*th International Conference on Science in Information Technology* (*ICSITech*), Yogyakarta, 23-24 October 2019, 83-88. https://doi.org/10.1109/icsitech46713.2019.8987530

[41] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., *et al.* (2001) Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, **17**, 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[42] Gad, I., Elmezain, M., Alwateer, M.M., Almaliki, M., Elmarhomy, G. and Atlam, E. (2023) Breast Cancer Diagnosis Using a Machine Learning Model and Swarm Intelligence Approach. 2023 1*st International Conference on Advanced Innovations in Smart Cities* (*ICAISC*), Jeddah, 23-25 January 2023, 1-5. https://doi.org/10.1109/icaisc56366.2023.10085393

[43] Mallinson, H. and Gammerman, A. (2003) Imputation Using Support Vector Machines. University of London Egham, UK: Department of Computer Science Royal Holloway.

[44] Noor, T.H., Almars, A., Gad, I., Atlam, E. and Elmezain, M. (2022) Spatial Impressions Monitoring during COVID-19 Pandemic Using Machine Learning Techniques. *Computers*, **11**, Article 52. https://doi.org/10.3390/computers11040052

[45] Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P. (2008) Extracting and Composing Robust Features with Denoising Autoencoders. *Proceedings of the* 25*th international conference on Machine learning—ICML*'08, Helsinki, 5-9 July 2008, 1096-1103. https://doi.org/10.1145/1390156.1390294

[46] Noor, T.H., Almars, A.M., Atlam, E. and Noor, A. (2022) Deep Learning Model for Predicting Consumers' Interests of IoT Recommendation System. *International Journal of Advanced Computer Science and Applications*, **13**, 161-170. https://doi.org/10.14569/ijacsa.2022.0131022

[47] Elmezain, M., Malki, A., Gad, I. and Atlam, E. (2022) Hybrid Deep Learning Model-Based Prediction of Images Related to Cyberbullying. *International Journal of Applied Mathematics and Computer Science*, **32**, 323-334. https://doi.org/10.34768/amcs-2022-0024

[48] Shahbazian, R. and Trubitsyna, I. (2022) DEGAIN: Generative-Adversarial-Network-Based Missing Data Imputation. *Information*, **13**, Article 575.

https://doi.org/10.3390/info13120575

[49] Malki, A., Atlam, E. and Gad, I. (2022) Machine Learning Approach of Detecting Anomalies and Forecasting Time-Series of IoT Devices. *Alexandria Engineering Journal*, **61**, 8973-8986. https://doi.org/10.1016/j.aej.2022.02.038

[50] Yoon, J., Jordon, J. and van der Schaar, M. (2018) GAIN: Missing Data Imputation Using Generative Adversarial Nets, *Proceedings of the* 35*th International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 5689-5698. https://proceedings.mlr.press/v80/yoon18a.html

[51] Chhabra, G. (2023) Comparison of Imputation Methods for Univariate Time Series. *International Journal on Recent and Innovation Trends in Computing and Communication*, **11**, 286-292. https://doi.org/10.17762/ijritcc.v11i2s.6148

[52] Malki, A., Atlam, E., Hassanien, A.E., Ewis, A., Dagnew, G. and Gad, I. (2022) SARIMA Model-Based Forecasting Required Number of COVID-19 Vaccines Globally and Empirical Analysis of Peoples' View towards the Vaccines. *Alexandria Engineering Journal*, **61**, 12091-12110. https://doi.org/10.1016/j.aej.2022.05.051

[53] Collins, L.M., Schafer, J.L. and Kam, C. (2001) A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, **6**, 330-351. https://doi.org/10.1037//1082-989x.6.4.330-351

[54] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O. (2021) A Survey on Missing Data in Machine Learning. *Journal of Big Data*, **8**, Article No. 140. https://doi.org/10.1186/s40537-021-00516-9

[55] Gustems-Carnicer, J. and Calderón, C. (2012) Coping Strategies and Psychological Well-Being among Teacher Education Students. *European Journal of Psychology of Education*, **28**, 1127-1140. https://doi.org/10.1007/s10212-012-0158-x

[56] Atlam, E., Masud, M., Rokaya, M., Meshref, H., Gad, I. and Almars, A.M. (2024) EASDM: Explainable Autism Spectrum Disorder Model Based on Deep Learning. *Journal of Disability Research*, **3**, 1-15. https://doi.org/10.57197/jdr-2024-0003

[57] Masud, M., Almars, A.M., Rokaya, M.B., Meshref, H., Gad, I. and Atlam, E. (2024) A Novel Light-Weight Convolutional Neural Network Model to Predict Alzheimer's Disease Applying Weighted Loss Function. *Journal of Disability Research*, **3**, 1-10. https://doi.org/10.57197/jdr-2024-0042

[58] McMahon, P., Zhang, T. and Dwight, R.A. (2020) Approaches to Dealing with Missing Data in Railway Asset Management. *IEEE Access*, **8**, 48177-48194. https://doi.org/10.1109/access.2020.2978902

[59] Bennin, K.E., Tahir, A., MacDonell, S.G. and Börstler, J. (2021) An Empirical Study on the Effectiveness of Data Resampling Approaches for Cross-Project Software Defect Prediction. *IET Software*, **16**, 185-199. https://doi.org/10.1049/sfw2.12052

[60] Hsu, C.C. and Sandford, B.A. (2019) Minimizing Non-Response in the Delphi Process: How to Respond to Non-Response. *Practical Assessment, Research, and Evaluation*, **12**, 17.

[61] Carpenter, J.R., Roger, J.H. and Kenward, M.G. (2013) Analysis of Longitudinal Trials with Protocol Deviation: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation. *Journal of Biopharmaceutical Statistics*, **23**, 1352-1371. https://doi.org/10.1080/10543406.2013.834911

[62] Adnan, F.A., Jamaludin, K.R., Wan Muhamad, W.Z.A. and Miskon, S. (2022) A Review of the Current Publication Trends on Missing Data Imputation over Three Decades: Direction and Future Research. *Neural Computing and Applications*, **34**, 18325-18340. https://doi.org/10.1007/s00521-022-07702-7

[63] Pedersen, A., Mikkelsen, E., Cronin-Fenton, D., Kristensen, N., Pham, T.M., Pedersen, L., *et al.* (2017) Missing Data and Multiple Imputation in Clinical Epidemiological Research. *Clinical Epidemiology*, **9**, 157-166. https://doi.org/10.2147/clep.s129785

[64] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, **37**, 50-60. https://doi.org/10.1109/msp.2020.2975749

[65] Singh, P., Singh, M.K., Singh, R. and Singh, N. (2022) Federated Learning: Challenges, Methods, and Future Directions. In: Yadav, S.P., Bhati, B.S., Mahato, D.P. and Kumar, S., Eds., *Federated Learning for IoT Applications*, Springer International Publishing, 199-214. https://doi.org/10.1007/978-3-030-85559-8_13

[66] Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J.M., Nowozin, S. and Zhang, C. (2018) Eddi: Efficient Dynamic Discovery of High-Value Information with Partial Vae. arXiv: 1809.11142.

[67] Sultana, Z., Akter, S. and Yeasmin, A. (2022) Transfer Learning Approach Applied to Data Imputation. University of Liberal Arts Bangladesh. https://doi.org/10.31219/osf.io/jd5th

[68] Lyu, L., Hu, Y., Wang, N., Zhou, X. and Fang, M. (2022) Application of Deep Learning and Transfer Learning in Continuous Missing Value Imputation of Water Quality Data. 2022 *IEEE 8th International Conference on Computer and Communications* (*ICCC*), Chengdu, 9-12 December 2022, 1211-1216. https://doi.org/10.1109/iccc56324.2022.10065657

[69] Anagnostopoulos, I. (2018) Fintech and Regtech: Impact on Regulators and Banks. *Journal of Economics and Business*, **100**, 7-25. https://doi.org/10.1016/j.jeconbus.2018.07.003

[70] Gupta, M. and Gupta, B. (2020) A New Scalable Approach for Missing Value Imputation in High-Throughput Microarray Data on Apache Spark. *International Journal of Data Mining and Bioinformatics*, **23**, 79-100. https://doi.org/10.1504/ijdmb.2020.105438

[71] Petrozziello, A., Jordanov, I. and Sommeregger, C. (2018) Distributed Neural Networks for Missing Big Data Imputation. 2018 *International Joint Conference on Neural Networks* (*IJCNN*), Rio de Janeiro, 8-13 July 2018, 1-8. https://doi.org/10.1109/ijcnn.2018.8489488

[72] Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H. (2018) Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, **169**, 866-872. https://doi.org/10.7326/m18-1990

[73] Chandler, R.K., Fletcher, B.W. and Volkow, N.D. (2009) Treating Drug Abuse and Addiction in the Criminal Justice System. *JAMA*, **301**, 183-190. https://doi.org/10.1001/jama.2008.976

[74] De Pau, M., Vruggink, R., Vandevelde, S. and Vander Laenen, F. (2023) Culturally Sensitive Forensic Mental Healthcare for Racialized People Labeled as Not Criminally Responsible: A Scoping Review. *International Journal of Forensic Mental Health*, **22**, 276-288. https://doi.org/10.1080/14999013.2023.2167891