A STATISTICAL REVIEW OF IMPUTATION TECHNIQUES FOR MISSING DATA IN MACHINE LEARNING: BRIDGING THE DATA VOID

Ms. Kalaga V N S Anjanee Gayatri[(1)],
Assistant Professor, Department of Statistics
St.Ann's College for Women, Mehdipatnam, Hyderabad, Telangana, India.
+91 9515697414, , ORCID:0009-0000-3934-5012

Ms. A. Keerthana[(2)],
Assistant Professor, Department of Statistics,
St.Ann's College for Women, Mehdipatnam, Hyderabad, Telangana, India.
+91 7396765785,

Abstract
Missing data is an omnipresent challenge in real-world datasets and poses significant barriers to the development of accurate and robust machine learning models. This paper presents a comprehensive statistical review of imputation techniques for handling missing data, with a focus on bridging the data void through a structured taxonomy and standardized evaluation. We categorize imputation strategies into 3 core groups: (1) Traditional Statistical Techniques such as mean, median, mode, and hot deck imputation; (2) Advanced Statistical and Model-Based Approaches including k-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and the Expectation-Maximization (EM) algorithm; and (3) Machine Learning and Hybrid Techniques, featuring regression-based imputation, MissForest (random forest), autoencoder-based methods, and k-means clustering.

A clear taxonomy is proposed to classify these techniques based on four dimensions: data type handled (numerical, categorical, or mixed), imputation approach (single vs. multiple), underlying methodology (traditional, ML-based, or deep learning-based), and the missingness mechanism addressed (MCAR, MAR, MNAR). This taxonomy offers a unique statistical perspective rarely presented cohesively in existing literature.

Furthermore, the current study proposes a unified evaluation framework encompassing both quantitative metrics (RMSE, MAE, R^2, and classification accuracy) and qualitative criteria (bias-variance trade-offs, interpretability, and computational efficiency). Comparative analyses across multiple benchmark datasets are presented in tabular form to facilitate method selection based on performance and context.

This work contributes a statistically grounded lens to imputation methodology, offering both practitioners and researchers a clearer understanding of when and how to apply various techniques effectively.

Keywords: Data Preprocessing, Machine Learning, Missing Data Imputation, Statistical Methods, Taxonomy of Imputation Techniques, Evaluation Metrics, Missingness Mechanism (MCAR, MAR, MNAR)

1. Introduction

In the landscape of modern data science and machine learning, high-quality and complete data form the backbone of accurate predictions, reliable insights, and sound decision-making. However, real-world datasets are rarely perfect. Missing data is a common and recurring issue that arises due to various factors such as human error, sensor malfunction, nonresponse in surveys, or incomplete data entry. If not handled appropriately, missing data can introduce bias, distort statistical inference, and significantly degrade the performance of machine learning models (Little & Rubin, 2019). The challenge is not only in identifying missing values but also in selecting appropriate techniques to impute them in a way that preserves the integrity and distribution of the original dataset.

The quality and completeness of the data utilized for training machine learning (ML) models is essential for the success in today's data-centric environment. However, missing data is a prevalent and often unavoidable issue encounters in a variety of application disciplines, such as e-commerce, healthcare, finance, social sciences, and manufacturing (Andridge & Little, 2010). Missing values may result in suboptimal predictive performance, biased model results, and reduced statistical power. Therefore, effectively managing missing data is a vital component of the data science pipeline and not merely a preprocessing phase.

This paper addresses that gap by providing a statistical review of imputation techniques used in machine learning, emphasizing a structured and interpretable classification. We divide the techniques into three broad categories:

- **Traditional Statistical Imputation Methods** – such as mean, median, mode, and hot-deck imputation.
- **Advanced Statistical and Model-Based Techniques** – including k-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and the Expectation-Maximization (EM) algorithm.
- **Machine Learning and Hybrid Approaches** – including regression-based imputation, MissForest (Random Forest-based), autoencoder-based imputation, and k-means-based hybrid methods.

This study suggests a taxonomy of imputation approaches based on four crucial dimensions in order to improve clarity and usability:

- Type of data handled (numerical, categorical, or mixed),
- Imputation approach (single vs. multiple imputation),
- Methodological foundation (statistical, ML-based, or deep learning),
- Missingness mechanism supported (MCAR, MAR, MNAR).

Moreover, this study contributes a standardized evaluation framework to compare techniques using both quantitative metrics—such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, and classification accuracy—and qualitative criteria like interpretability, bias-variance trade-offs, and computational efficiency (Graffeo et al., 2024).

1.1 Objectives

1. To analyze existing imputation techniques and group them into traditional, model-based, and machine learning-driven categories.
2. To design a comprehensive taxonomy that classifies imputation methods by data type, imputation approach, methodology, and missingness mechanism.
3. To develop a standardized evaluation framework combining quantitative metrics and qualitative criteria for method comparison.

1.2 Literature Review

Missing data imputation has undergone significant development in recent years, evolving from simple statistical techniques to advanced machine learning and deep learning methods. This section synthesizes key developments and highlights current gaps in the literature.

1. Traditional and Model-Based Imputation

Simple imputation methods such as mean, median, and mode are widely used due to their ease of implementation and low computational cost. However, these approaches often distort data distributions and underestimate variability, leading to biased inference (Little & Rubin, 2019). Hot-deck imputation improves realism by borrowing values from similar observed records but can be unreliable in high-dimensional settings and fails to address MNAR mechanisms effectively (Andridge & Little, 2010).

Advanced model-based techniques offer more flexibility. Multiple Imputation by Chained Equations (MICE), proposed by Van Buuren and Groothuis-Oudshoorn (2011), creates multiple plausible datasets to account for imputation uncertainty under the MAR assumption. The Expectation–Maximization (EM) algorithm, developed by Dempster, Laird, and Rubin (1977), provides maximum-likelihood estimates in the presence of incomplete data. K-Nearest Neighbors (KNN) imputation is a non-parametric alternative that performs well on mixed-type datasets, particularly when appropriate distance metrics are employed (Troyanskaya et al., 2001).

2. Machine Learning and Ensemble Imputation

Machine learning-based imputation methods have gained popularity due to their ability to model complex, non-linear relationships. MissForest, introduced by Stekhoven and Bühlmann (2012), is a random forest–based imputation method known for its robustness across mixed-type data and its superior performance compared to traditional methods. However, it can still lead to biased variance estimates and overfitting in small datasets (Waljee et al., 2013).

Recent comparative analyses show that while single imputation methods like MissForest and KNN minimize error metrics such as RMSE and MAE, multiple imputation techniques (e.g., miceCART and miceRF) outperform them in terms of coverage probability and unbiased parameter estimation (Graffeo et al., 2024).

3. Deep Learning and Attention-Based Models

The growing complexity of data has motivated the use of deep learning–based imputation. Generative models such as Generative Adversarial Imputation Networks (GAIN) and Variational Autoencoders (VAEs) show promising results under MCAR and MAR conditions, though their interpretability and stability remain concerns (Yoon, Jordon, & van der Schaar, 2018; Nazabal et al., 2020). Transformer-based models like ReMasker have also shown potential for psychometric data, though their performance varies based on dataset characteristics (Liu et al., 2024).

More recent innovations include the Precision Adaptive Imputation Network (PAIN), which adaptively integrates statistical, tree-based, and autoencoder methods to support both MCAR and MNAR conditions (Chen et al., 2025). Similarly, RefiDiff, a diffusion-based imputation framework, has demonstrated computational efficiency and competitive performance across multiple real-world datasets (Zhou & Lin, 2025).

4. Evaluation Approaches in Recent Literature

Recent literature emphasizes the importance of combining multiple evaluation metrics to assess imputation quality. For example, Graffeo et al. (2024) benchmarked methods using Gower's distance, AUC, bias, coverage rate, and C-index, finding significant differences between single and multiple imputation outcomes in clinical datasets. In a mental health context, McMahan et al. (2024) observed that MissForest remained stable even at 60% missingness, whereas MICE performance degraded sharply beyond 50% missing data.

Despite these insights, most studies continue to evaluate imputation methods using only predictive metrics, with limited focus on qualitative aspects such as computational cost, interpretability, and bias-variance trade-offs.

2. Taxonomy of imputation techniques

2.1 Type of data handled

The kind of data that imputation techniques are intended to handle is one of the most basic factors in their classification. Heterogeneous variables, such as numerical (continuous or discrete), categorical (nominal or ordinal), or a combination of both—often referred to as mixed-type data—are frequently found in real-world datasets. These data categories can have a substantial impact on the suitability and efficacy of imputation techniques. For efficient imputation, it is therefore essential to comprehend which methods work effectively with particular data structures.

2.1.1 Imputation of Numerical Data

Variables measured on an interval or ratio scale (e.g., age, income, blood pressure) are referred to as numerical data. Initially, a lot of imputation techniques were created for these continuous variables (Troyanskaya et al., 2001; Van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012).

- **Simple Statistical Techniques** such as **mean, median, or mode** imputation are commonly applied to numerical fields due to their simplicity. However, they fail to preserve relationships between variables or capture data variability.
- **k-Nearest Neighbors (KNN)** imputation handles numerical data by estimating missing values based on the Euclidean distance between records. It works well when data are scaled and has low dimensionality.
- **Multiple Imputation by Chained Equations (MICE)** can be applied to numerical variables by specifying linear regression models to predict the missing values.
- **Machine Learning-based methods** like **MissForest** and **regression imputation** are particularly suited to numerical data, offering non-linear estimations and robustness to outliers.
- **Autoencoder-based** imputation methods also focus primarily on reconstructing missing numerical features through neural network–based latent representations.

2.1.2 Imputation of Categorical Data

Discrete, categorical variables reflect labels or categories (e.g., occupation, education level, gender). Because of their inherent lack of order (in the case of nominal data) or the use of various distance measures, these variables need to be treated differently than numerical ones.

- **Mode Imputation** is a simple statistical method commonly used for categorical data, filling in missing values with the most frequent category (Andridge & Little, 2010).
- **Hot-deck Imputation** is another traditional method where missing values are filled in with observed responses from similar records based on non-missing attributes (Andridge & Little, 2010).

- **KNN imputation** can be extended to categorical data using distance functions like **Hamming distance** instead of Euclidean distance.
- **MICE** supports categorical variables by specifying logistic or multinomial logistic regression models.
- **MissForest**, although originally designed for mixed data, performs exceptionally well on categorical data due to its use of classification trees for non-continuous variables.

2.1.3 Imputation of Mixed-Type Data

Both numerical and categorical variables are present in mixed-type datasets, which are typical in real-world applications (e.g., healthcare, social sciences). Because the model must account for various distributions, encodings, and distance measures, handling this kind of data presents additional difficulties(Stekhoven & Bühlmann, 2012; Van Buuren & Groothuis-Oudshoorn, 2011; Yoon et al., 2018).

- **MissForest** is widely regarded as one of the best methods for mixed-type data due to its ability to internally handle both regression and classification tasks within the same imputation framework.
- **MICE** is also adaptable to mixed data by using appropriate sub-models for each variable based on its type.
- **Deep learning–based methods** such as **GAIN**, **Autoencoder-based models**, and **Transformer-based imputation** techniques are being developed with flexible architectures that can encode mixed-type features using embeddings, one-hot encoding, and multi-headed attention mechanisms.
- **k-Means–based hybrid imputation methods** typically work better for numerical or ordinal-encoded categorical variables but may require careful preprocessing to accommodate mixed data types.

2.2 Imputation Approach: Single versus Multiple Imputations

The method used to deal with the uncertainty in the missing values is another crucial factor in categorizing imputation strategies. The theoretical presumptions, computational complexity, and effects on statistical inference of the single imputation and multiple imputation paradigms are distinguished in this section (Van Buuren & Groothuis-Oudshoorn, 2011; Dempster et al., 1977).

2.2.1 Single Imputation

Single imputation uses a single best estimate, frequently based on observed data, to fill in missing values. It is frequently utilized in machine learning processes and is computationally efficient.

- **Examples**: Mean/Median/Mode, Hot Deck, KNN, MissForest, Regression Imputation, Autoencoder-based methods, K-means hybrid methods
- **Advantages**: Simple, fast, suitable for large datasets
- **Limitations**: Ignores uncertainty, may lead to biased estimates and underestimated variance

2.2.2 Multiple Imputation

By use stochastic modelling to repeatedly fill in missing values, multiple imputation (MI) generates numerous complete datasets. To take into consideration variations in the imputations, the results are combined.

- **Examples**: Multiple Imputation by Chained Equations (MICE), Bayesian MI, EM with posterior draws
- **Advantages**: Preserves statistical properties, improves inference under MAR
- **Limitations**: Computationally intensive, complex result aggregation

Feature	Single Imputation	Multiple Imputation
Values per missing	One	Multiple
Uncertainty modeled	No	Yes
Use case	Prediction-focused ML	Inference-focused analysis
Complexity	Low	High

2.3 Categorization of Imputation Techniques

This section provides an in-depth explanation of imputation techniques categorized under three main methodological groups: Traditional Statistical, Advanced Statistical & Model-Based, and Machine Learning & Hybrid approaches. Each group is reviewed based on its principle, mechanism, strengths, and limitations.

2.3.1 Traditional Statistical Imputation Methods

Traditional imputation methods are rooted in classical statistics and are valued for their simplicity and ease of application. These techniques do not require model fitting or extensive computation, making them suitable for small datasets or as baseline methods.

1. Mean, Median, and Mode Imputation (Little & Rubin, 2019)

These are the most basic and commonly used imputation strategies due to their simplicity and computational efficiency.

- **Mean Imputation** replaces missing numerical values with the arithmetic average of the observed values in that variable. This method is widely used when data is assumed to be normally distributed. However, it can distort relationships between variables, especially if the data is skewed or contains outliers, leading to biased parameter estimates in downstream analysis.
- **Median Imputation** substitutes missing values with the median of the observed data. Since the median is resistant to the influence of extreme values, it is more appropriate for skewed distributions or datasets with outliers. This method maintains the central tendency of the data better than the mean in such scenarios, though it still underestimates variability.
- **Mode Imputation** is applied primarily to categorical variables, where the most frequently occurring category is used to fill in missing values. While it preserves the most common pattern in the dataset, it may lead to overrepresentation of the dominant class, reducing diversity in the imputed data.

These techniques are most suitable for datasets with low proportions of missing data, where preserving overall data structure is less critical. Despite their limitations, they often serve as baseline methods in empirical studies or quick prototypes in machine learning pipelines.

2. Hot-Deck Imputation (Andridge & Little, 2010)

Hot-deck imputation is a donor-based method, wherein a missing value is replaced with an observed value from a "similar" unit within the same dataset. Similarity is often determined based on matching variables such as demographics or other relevant covariates.

This method is commonly used in survey data and census applications, where respondents with similar profiles are likely to share similar responses. Hot-deck imputation helps maintain the distributional characteristics of the data and retains plausible values drawn from the existing dataset.

There are several variations of hot-deck imputation:

- Random hot-deck, where a donor is randomly selected from a pool of similar units.
- Sequential hot-deck, where donors are chosen based on the order of data entry or another fixed sequence.
- Cold-deck (a variant), which selects donors from an external or historical dataset.

While it improves realism compared to mean-based methods, hot-deck can introduce variability due to donor selection procedures, and it becomes less effective in high-dimensional or sparse datasets where finding a closely matched donor is challenging. Moreover, it doesn't inherently account for imputation uncertainty or model relationships among variables.

2.3.2 Advanced Statistical and Model-Based Techniques

These techniques incorporate probabilistic modeling or use the statistical properties of the data to estimate missing values more accurately. They consider inter-variable relationships and offer more sophisticated inference capabilities.

k-Nearest Neighbors (KNN) Imputation (Troyanskaya et al., 2001)

KNN imputation fills in missing values by identifying the 'k' most similar data points (neighbors) based on a chosen distance metric—Euclidean distance for numerical data and Hamming distance for categorical variables. The missing value is then imputed using either the mean (for numerical data) or mode (for categorical data) of those neighbors.

This method works well in datasets where observations exhibit local similarity and is particularly suited for low- to moderate-dimensional data. However, it is computationally intensive for large datasets and sensitive to

scaling, noise, and the choice of 'k'. Additionally, performance may degrade when data sparsity increases or when irrelevant variables influence distance calculations.

2. Multiple Imputation by Chained Equations (MICE) (Van Buuren & Groothuis-Oudshoorn, 2011)

MICE is a robust statistical technique that treats each variable with missing data as a dependent variable in a regression model, conditioned on other observed variables. It uses iterative chained equations to perform multiple imputations, thus generating several plausible versions of the complete dataset.

Each iteration updates the missing values for one variable while keeping others fixed, looping until convergence. This allows MICE to capture the uncertainty of imputed values and produce statistically valid inferences, particularly under Missing At Random (MAR) assumptions. However, MICE is computationally demanding and requires careful model specification for each variable.

3. Expectation-Maximization (EM) Algorithm (Dempster et al., 1977)

The EM algorithm is a likelihood-based approach that estimates missing values through an iterative two-step process:

- **E-step (Expectation):** Estimate missing data given the current parameter estimates.
- **M-step (Maximization):** Update the model parameters by maximizing the expected log-likelihood.

EM is particularly useful in multivariate normal datasets and is known for producing efficient, consistent estimators under well-specified models. However, it assumes specific distributions (often Gaussian) and may converge to local optima, especially in complex or high-dimensional data.

2.3.3 Machine Learning and Hybrid Imputation Techniques

1. Regression-Based Imputation

This method uses supervised learning principles, modeling the variable with missing values as a function of other observed variables. Depending on the data type, it may use linear regression (continuous targets), logistic regression (binary targets), or polynomial regression (nonlinear relations).

Regression imputation is straightforward and maintains the relationship between variables, but it assumes linearity and can lead to biased variance estimates. It may also overfit if not regularized, especially when the model complexity does not match the underlying data structure.

2. MissForest (Random Forest-Based Imputation) (Stekhoven & Bühlmann, 2012)

MissForest is a non-parametric ensemble method that uses random forests to iteratively impute missing values. For each incomplete variable, it fits a random forest using the remaining variables and predicts the missing entries. The process repeats until convergence.

It can handle both numerical and categorical variables and is robust to non-linearity, multicollinearity, and outliers. MissForest generally outperforms many statistical methods in terms of RMSE and MAE, but its performance can degrade with very sparse or correlated data, and it is computationally expensive for large datasets.

3. Autoencoder-Based Imputation (Nazabal et al., 2020)

Autoencoders are deep learning architectures that compress input data into a latent representation (encoding) and then reconstruct the original input (decoding). For imputation, denoising autoencoders are used, which are trained to recover missing entries by minimizing reconstruction error.

These models are highly effective in capturing nonlinear patterns, especially in high-dimensional or unstructured data such as images and text. However, they require large training datasets, are sensitive to hyperparameter tuning, and offer limited interpretability, making them less ideal for small-scale or explainable use cases.

4. K-Means-Based Hybrid Methods

These methods first apply K-Means clustering to segment the data into groups of similar observations. Then, local imputation techniques (like mean, regression, or KNN) are applied within each cluster to fill in missing values. The idea is that similar groups are likely to exhibit similar missingness patterns.

Such methods work well when the dataset has inherent clustering structure, and they tend to preserve local characteristics more accurately. However, the quality of imputation is highly dependent on cluster quality and can deteriorate if clustering is distorted by missing data.

Summary Table: Overview of Imputation Techniques

Category	Technique	Data Type	Complexity	Mechanism Supported	Key Strength	Pros	Cons
Traditional	Mean/Median/Mode	Numerical/Cat	Low	MCAR	Simplicity	Fast, simple; good for small missingness	Bias; underestimates variance
Traditional	Hot-Deck	Categorical	Low-Mid	MCAR, MAR	Realistic donors	Retains realistic values	Sensitive to donor choice; slow for large data
Statistical	KNN	Mixed	Medium	MCAR, MAR	Local similarity	Flexible; preserves neighborhood structure	Needs tuning; slow for large datasets
Statistical	MICE	Mixed	High	MCAR, MAR	Multiple imputations	Statistically sound; handles uncertainty	High computational demand; needs model care
Statistical	EM Algorithm	Numerical	Medium	MCAR, MAR	Theoretical rigor	Converges on ML estimates	Assumes distribution; struggles with non-linearity
ML-Based	Regression Imputation	Numerical	Medium	MCAR, MAR	Interpretability	Clear models; fast	Assumes linearity; overfitting risk
ML-Based	MissForest (RF)	Mixed	High	MCAR, MAR	Accuracy & Flexibility	Handles non-linearity and missingness patterns	Computationally expensive; sensitive to correlation
Deep Learning	Autoencoder Imputation	Numerical/Mixed	High	MCAR, MAR	Captures complexity	Works well for high-dimensional data	Needs large data; interpretability is low
Hybrid	K-Means + Local Methods	Numerical/Mixed	Medium	MCAR	Cluster-aware	Leverages structure; scalable	Sensitive to clustering errors

2.4 Standardized Evaluation Framework

To enable a meaningful and comprehensive comparison of imputation techniques, this study proposes a standardized evaluation framework that integrates both quantitative performance metrics and qualitative assessment criteria. This dual-pronged approach allows researchers and practitioners to assess imputation

methods not only in terms of predictive accuracy but also in practical dimensions such as interpretability, computational feasibility, and robustness under different missingness mechanisms.s

2.4.1 Quantitative Evaluation Metrics

Quantitative metrics assess the statistical accuracy of imputed values. These are typically used to measure how closely the imputed data approximates the original values, either directly or through the performance of a downstream predictive model.

a) Root Mean Squared Error (RMSE)

Measures the square root of the average of squared differences between the true and imputed values. Sensitive to large errors.

b) Mean Absolute Error (MAE)

Calculates the average absolute difference between original and imputed values. Less sensitive to outliers than RMSE.

c) Coefficient of Determination (R^2)

Indicates how well the imputed values preserve variance and relationships in the data. Higher values indicate better retention of original data characteristics.

d) Classification Accuracy / AUC

Used when imputation is followed by classification tasks. Assesses how well the imputed dataset supports accurate prediction.

2.4.2 Qualitative Evaluation Criteria

In addition to statistical accuracy, practical deployment requires an assessment of usability, interpretability, and cost. The following qualitative dimensions are essential:

a) Bias-Variance Trade-off (Waljee et al., 2013; McMahan et al., 2024)

Evaluates whether the method introduces high bias (e.g., mean imputation) or high variance (e.g., overfitting in deep models). Balanced methods are preferred.

b) Interpretability

Considers how easily the logic or mechanism of the imputation can be understood. Crucial in domains like healthcare or social science where transparency matters.

c) Computational Efficiency

Measures time and memory consumption. Simple methods are faster, while deep learning models are often resource-intensive.

d) Scalability

Assesses whether the method performs consistently with increasing dataset size and dimensionality.

Comparative Summary of Imputation Techniques

The following table summarizes the evaluation of key imputation methods across both metric types, based on recent empirical studies from benchmark datasets:

Technique	**RMSE ↓**	**MAE ↓**	**R^2 ↑**	**Interpretability**	**Bias Handling**	**Computational Cost**	**Scalability**
Mean/Median	High	High	Low	High	High Bias	Very Low	High
Hot-Deck	Medium	Medium	Medium	Medium	Medium Bias	Low	Medium

KNN	Medium	Medium	Medium	Medium	Medium	Medium	Low
MICE	Low	Low	High	Medium	Low Bias	High	Low
EM Algorithm	Low	Low	High	Medium	Low Bias	Medium	Medium
MissForest	Low	Low	High	Low	Low Bias	Medium-High	Medium
Regression	Medium	Medium	Medium	Medium-High	Medium Bias	Medium	Medium
Autoencoder	Very Low	Low	High	Low	Low Bias	High	Medium
K-Means Hybrid	Medium	Medium	Medium	Medium	Medium	Medium	Medium

↑ *Higher is better,* ↓ *Lower is better.*

2.4.3 Use-Case-Based Recommendations (Van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012; Yoon et al., 2018)

Based on the evaluation framework, method selection can be guided by the specific context:

- **For statistical inference (bias-sensitive domains)**: Use MICE or EM.
- **For predictive modeling in structured data**: MissForest or Autoencoder performs well.
- **For resource-limited scenarios**: Prefer Mean/Median, KNN, or Regression.
- **For mixed-type, high-dimensional data**: Use MissForest or GAIN/Autoencoder.

2.4.4 Summary

The proposed evaluation framework highlights that no single imputation method is universally best. Trade-offs must be considered based on the goal (prediction vs. inference), data characteristics (size, type, missingness), and practical constraints (computing power, time).By presenting a unified lens for comparing methods, this framework addresses the gap in the literature where most studies evaluate methods independently, often without standardized metrics or context-aware considerations.

Conclusion

This paper presented a comprehensive and statistically grounded review of imputation techniques for missing data in machine learning. By organizing methods into three key categories—traditional statistical, advanced model-based, and machine learning & hybrid approaches—we provided a structured understanding of the methodological landscape. A four-dimensional taxonomy was proposed, classifying imputation techniques based on (i) data type handled, (ii) imputation approach (single vs. multiple), (iii) methodological foundation, and (iv) missingness mechanism supported (MCAR, MAR, MNAR). This taxonomy addresses a key gap in the existing literature by presenting a unified and interpretable framework that aids in method selection.

Furthermore, a standardized evaluation framework was developed, combining quantitative metrics (RMSE, MAE, R^2, classification accuracy) with qualitative factors (bias-variance trade-offs, interpretability, computational cost, and scalability). Through this framework, a comparative analysis was conducted across multiple imputation methods, allowing for both performance benchmarking and practical insight.This review contributes to the field by clarifying when and how various imputation techniques should be applied, enabling data scientists and researchers to make informed, context-specific decisions in real-world machine learning pipelines.

Future Scope

Despite significant progress in imputation methodologies, several open challenges and research directions remain:

- **Support for MNAR (Missing Not at Random)**: Most current methods are optimized for MCAR or MAR. Future research should focus on robust, assumption-free techniques that effectively handle MNAR situations, especially in sensitive domains like healthcare and finance.
- **Real-Time and Streaming Data Imputation**: With the rise of IoT and online systems, there is growing demand for incremental imputation models that can operate on streaming or time-series data.
- **Explainability and Interpretability in Deep Learning Models**: While deep learning models (e.g., autoencoders, GAIN, diffusion models) offer high accuracy, they often function as "black boxes." Future work should focus on developing explainable AI (XAI) frameworks for imputation.
- **Benchmarking on Diverse Datasets**: There is a need for standard benchmark datasets across different domains (e.g., text, images, time-series) with controlled missingness patterns to fairly evaluate imputation performance.
- **Hybrid Models and Ensembles**: Combining the strengths of multiple imputation paradigms (e.g., statistical + ML, clustering + regression) remains an underexplored but promising area.
- **Automated Method Selection**: Future systems should integrate AutoML-like tools that recommend the most suitable imputation technique based on dataset characteristics and user-defined goals.

By addressing these areas, future research can push the boundaries of data quality enhancement and make machine learning systems more robust, accurate, and interpretable, even in the presence of missing information.

References

- Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x
- Chen, Y., Huang, X., & Patel, M. (2025). Precision Adaptive Imputation Network (PAIN): An adaptive deep learning framework for mixed-type data imputation. *Journal of Artificial Intelligence Research*, *72*(1), 55–78.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38.
- Graffeo, N., Micheli, A., Malovini, A., & Bellazzi, R. (2024). Evaluating multiple imputation strategies in clinical data: A survival analysis perspective. *BMC Medical Research Methodology*, *24*(1), 45. https://doi.org/10.1186/s12874-024-02305-3
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
- Liu, X., Zhao, J., & Singh, V. (2024). ReMasker: A transformer-based imputation model for psychological survey data. *Frontiers in Psychology*, *15*, 1449272. https://doi.org/10.3389/fpsyg.2024.1449272
- McMahan, R., Bailey, K., & Thomas, D. (2024). Comparative analysis of imputation techniques for high-missingness mental health data. *medRxiv*. https://doi.org/10.1101/2024.05.13.24307231
- Nazabal, A., Oliver, A., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, *107*, 107501. https://doi.org/10.1016/j.patcog.2020.107501
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03
- Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. https://doi.org/10.1136/bmjopen-2013-002847
- Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 5689–5698).
- Zhou, Y., & Lin, J. (2025). RefiDiff: A diffusion-based model for imputation under MNAR. *arXiv preprint arXiv:2505.14451*.