

# A survey of missing data imputation techniques: statistical methods, machine learning models, and GAN-based approaches

Rifaa Sadeh, Ahmed Mohameden, Mohamed Lemine Salihi, Mohamedade Farouk Nanne

Scientific Computing, Computer Science and Data Science, Department of Computer Science, Faculty of Science and Technology,  
University of Nouakchott, Nouakchott, Mauritania

## Article Info

### Article history:

Received Jun 8, 2024

Revised Jun 11, 2025

Accepted Jul 10, 2025

### Keywords:

Data imputation

Generative adversarial networks

Machine learning

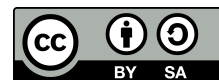
Missing data

Statistical methods

## ABSTRACT

Efficiently addressing missing data is critical in data analysis across diverse domains. This study evaluates traditional statistical, machine learning, and generative adversarial network (GAN)-based imputation methods, emphasizing their strengths, limitations, and applicability to different data types and missing data mechanisms (missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR)). GAN-based models, including generative adversarial imputation network (GAIN), view imputation generative adversarial network (VIGAN), and SolarGAN, are highlighted for their adaptability and effectiveness in handling complex datasets, such as images and time series. Despite challenges like computational demands, GANs outperform conventional methods in capturing non-linear dependencies. Future work includes optimizing GAN architectures for broader data types and exploring hybrid models to enhance imputation accuracy and scalability in real-world applications.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Rifaa Sadeh

Scientific Computing, Computer Science and Data Science, Department of Computer Science

Faculty of Science and Technology, University of Nouakchott

Nouakchott, Mauritania

Email: rifasadeh@gmail.com

## 1. INTRODUCTION

Missing data is a pervasive challenge that affects nearly every scientific discipline, from medicine [1] to geology [2], energy [3] and environmental sciences [4]. Rubin [5] defined missing data as unobserved values that could yield critical insights if available. These gaps introduce biases, distort analysis, and reduce the effectiveness of algorithms, ultimately impairing decision-making processes.

The origins of missing data are diverse, arising from incomplete data collection, recording errors, or hardware malfunctions [5]. These gaps skew results and misrepresent the studied population [6], creating a need for robust and scalable solutions to ensure reliable research outcomes. Addressing missing data has proven to be a multifaceted problem, requiring methods that vary depending on the type and complexity of the dataset. Initial approaches, such as listwise deletion, were simple but often discarded valuable information along with the missing data [7]. Over time, more sophisticated imputation techniques emerged, including statistical methods, machine learning algorithms, and deep learning models. Among these, generative adversarial networks (GANs) have gained prominence for their ability to model complex data distributions and address non-linear dependencies effectively. Despite their potential, implementing GANs for data imputation comes

with challenges, including: i) high computational costs due to complex training processes; ii) sensitivity to hyperparameter tuning, which affects model stability; and iii) risk of overfitting, particularly when handling small datasets.

This paper provides a comprehensive review of missing data imputation methods. We analyze traditional statistical approaches, machine learning techniques, and deep learning models, with a particular focus on GAN-based imputation. Our findings reveal that while GANs outperform traditional methods in handling complex datasets, their deployment requires careful balancing of model complexity and computational efficiency. We also propose future research directions, including: i) the integration of hybrid models combining statistical techniques with GANs; ii) optimization of GAN architectures for imputation tasks; and iii) application of these techniques to real-world datasets in fields such as healthcare, energy, and environmental science. By addressing these challenges and exploring innovative solutions, this work aims to contribute to the growing body of knowledge in data imputation, enabling researchers and practitioners to better handle missing data scenarios.

The remainder of this article is structured as follows: section 2 introduces the methodology and criteria for evaluating imputation methods. Section 3 presents a comparative analysis of different approaches. Section 4 discusses the implications of the results, including ethical considerations related to imputation in sensitive domains. Section 5 concludes with key findings and recommendations for future research.

## 2. MISSING DATA MECHANISMS AND TYPES OF VARIABLES

Handling missing data is critical for ensuring the reliability of statistical analyses. Understanding the mechanisms underlying missing data and the types of variables involved is fundamental for selecting appropriate imputation techniques. This section explores the categories of missing completely at random (MCAR), missing not at random (MNAR), and missing at random (MAR), alongside a classification of statistical variables and imputation approaches.

### 2.1. Missing data categories

Missing data can be classified into three distinct categories: MCAR, MNAR and MAR [5].

- MCAR: data is missing randomly, unrelated to observed or unobserved variables.  
Example: pixels missing in radiological images due to random noise or technical errors, such as sensor malfunction.
- MAR: missingness depends on observed variables.  
Example: crop yield data missing in regions with extreme weather conditions, where meteorological data is recorded.
- MNAR: missingness depends on unobserved variables [8].  
Example: fetal position affects the visibility of genital organs during an ultrasound, leading to gender data being systematically missing when the fetus is positioned laterally or with crossed legs.

Figure 1 provides an illustration. Table 1 summarizes the criteria distinguishing these categories. MCAR is ignorable, while MAR and MNAR require advanced techniques to mitigate bias.  $P(M = 1|Y_o, Y_m, \psi)$  defines the probability of the missing data mechanism, where  $\psi$  represents the set of parameters of the imputation model. When data is MNAR, the probability of the mechanism cannot be defined because it depends on one or more unmeasured parameters, i.e., unobserved variables.

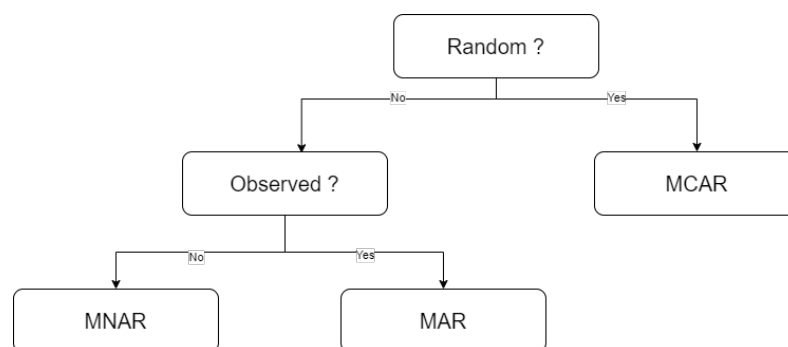


Figure 1. Missing data mechanisms

Table 1. Comparison of missing data mechanisms

Criterion	MAR	MNAR	MCAR
Random	No	No	Yes
Ignorable	It depends	No	Yes
Dependency	Observed variable	Unobserved variable	None
$P(M = 1 Y_o, Y_m, \psi)$	$P(M = 1 Y_o, \psi)$	Undefined	$P(M = 1, \psi)$

## 2.2. Imputation approaches

Imputation methods are categorized based on variable relationships:

- Single vs. multiple imputation: single imputation replaces a missing value with one estimate, while multiple imputation generates several plausible values [9].
- Univariate vs. multivariate: univariate imputation considers only the target variable, whereas multivariate imputation incorporates relationships between variables [10].

Multivariate methods are preferable for datasets with strong interdependencies as they support both single and multiple imputations, as shown in Table 2.

Table 2. Comparison of imputation types

Criterion	Approach	
	Univariate	Multivariate
Replacement	1	$m$
Correlation	×	✓
Single Imputation	×	✓
Multiple Imputations	✓	✓

## 2.3. Types of variables

Statistical variables are classified as: i) quantitative (e.g., continuous: salary, discrete: age). ii) qualitative (e.g., nominal: marital status, ordinal: satisfaction level) [11]. Misinterpretations arise when qualitative variables are numerically encoded (e.g., zip codes), as their mean has no significance. Figure 2 provides an overview.

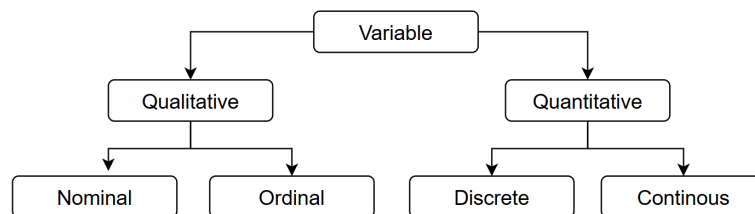


Figure 2. Types of statistical variables

## 3. IMPUTATION METHODS

Managing missing data is crucial across various fields to ensure the accuracy of analyses and predictive models. This section reviews several imputation techniques, ranging from traditional statistical methods to advanced machine learning and deep learning approaches. Each method's strengths and limitations are discussed, along with their suitability for different data types and contexts.

### 3.1. Statistical methods

Statistical methods are foundational for imputation. Key approaches include similarity-based methods, observation-based methods, measures of central tendency, and multivariate imputation by chained equations (MICE).

#### 3.1.1. Similarity-based methods

The hot-deck method replaces missing values with those from similar individuals. The cold-deck method uses values from external sources. This is applied when there are not enough similar data points [12], [13].

### 3.1.2. Observation-based methods

Methods like last observation carried forward (LOCF), baseline observation carried forward (BOCF), worst observation carried forward (WOCF), and next observation carried backward (NOCB) are commonly used for longitudinal data. These methods replace missing values based on temporal patterns. They rely on the assumption that nearby observations carry meaningful information [14]-[17].

### 3.1.3. Measures of central tendency

The objective of central tendency measures is to summarize, in a single value, the elements of a variable in a dataset. The most commonly used central tendency measures are the mean [18], the median [19], and the mode [20]. Indeed, there are various means [21], such as the arithmetic, quadratic, harmonic, geometric, weighted, and truncated means. Here, we illustrate the arithmetic mean, where the imputation involves replacing the missing values of a variable with the sum of its known values, divided by the total number of values:

$$\forall i \in \{1, 2, \dots, p\}, \bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} \mid y_{ij} \notin Y_m$$

The arithmetic mean is only applicable to quantitative variables, especially continuous ones. However, it can also be used for discrete variables, in which case the result will be rounded to the nearest integer.

The median is the value that divides the elements of an observed variable into two equal parts. After sorting the values of the target observed variable in ascending order, imputation by the median involves replacing the missing values of a variable with the middle value when the number of observations  $n$  is odd, or the average of the two middle observations if  $n$  is even:

$$\forall i \in \{1, 2, \dots, p\}, \tilde{y}_i = \begin{cases} y_{\frac{n+1}{2}} & \text{if } n \equiv 0 \pmod{2} \\ y_{\frac{n}{2}} & \text{if } n \equiv 1 \pmod{2} \end{cases}$$

In addition to the classical median, there are other ways [22] to calculate a measure of central position, such as the weighted median, the geometric median, and the absolute median deviation. Imputation by mode replaces missing data with the most frequent value of the target variable:

$$\forall i \in \{1, 2, \dots, p\} \quad \exists^* j \in \{1, 2, \dots, n\} \quad \text{such that} \quad \hat{y}_i = \operatorname{argmax}_{y_{ij}} P(Y = y_{ij})$$

Although the mode can be calculated for both numerical and categorical variables, in practice, it is commonly used only for nominal variables as they do not have other central tendency measures.

### 3.1.4. Multivariate imputation by chained equations

MICE is an iterative approach that imputes missing data using regression models. Each missing value is predicted using a regression model based on other variables in the dataset. The algorithm iterates until the imputed values converge [23], [24].

## 3.2. Machine learning methods

Machine learning methods offer advantages over traditional statistical approaches, particularly in handling large and complex datasets [25]. This section reviews four popular machine learning models for data imputation: linear regression, logistic regression, k-nearest neighbors (KNN), and decision trees.

### 3.2.1. Regression

Regression models estimate relationships between the target and observed variables. We focus on linear and logistic regression [26].

- Linear regression models aim to capture a proportional trend between inputs and outcomes. It operates by applying the least squares method to reduce the gap between actual observations and model predictions.

$$y = \alpha x + \beta + \epsilon \quad (1)$$

Here,  $\alpha$  is the coefficients of the regression line and  $\beta$  originally ordered, and  $\epsilon$  is the error term, representing the unexplained deviation or variance by the linear relationship between the observed value  $y$  and the predicted value  $\alpha + \beta x$ .

- Logistic regression: used for binary classification, it models the probability of the target variable being 1 using a logistic function:

$$p = \frac{1}{1 + e^{-z}} \quad (2)$$

Here,  $p$  is the probability that the target variable is 1, and  $z$  is the linear function in the form:

$$z = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (3)$$

Where  $b_0, b_1, b_2, \dots, b_n$  are the regression coefficients, and  $x_1, x_2, \dots, x_n$  are the observed variables.

### 3.2.2. K-nearest neighbors

The basic idea of the KNN is to find the  $k$ -nearest neighbors of the individual with missing data [27]. This algorithm requires two parameters, namely, the value of  $k$  and the similarity metric between individuals. The similarity is calculated using a distance measure such as the Euclidean distance, the Manhattan distance, and the Minkowski distance.

### 3.2.3. Decision trees

Decision trees partition data into subsets based on feature values to predict missing values. Random forests, an ensemble of multiple decision trees trained on different subsets, enhance robustness and reduce overfitting. MissForest [28], a widely used variant, begins with naive imputations and iteratively refines predictions via random forests. These methods are more flexible than traditional statistical approaches but may require careful tuning for high-dimensional or sparse datasets.

## 3.3. Deep learning methods

Deep learning models offer two major advantages over traditional machine learning models. Firstly, traditional methods often require manual selection of relevant features or variables for training the imputation model. In contrast, deep learning models use neural networks to automatically learn these features from raw data. This "automation" occurs during the learning phase, where biases and weights in each layer of the neural network are adjusted to better capture the underlying patterns in the data. The second advantage is the versatility of neural networks, which makes them easily adaptable to various scenarios, including the 12 cases illustrated in Table 3. Neural networks can model complex, non-linear relationships, making them particularly effective for imputing data with intricate patterns.

Table 3. Overview of methods for imputing missing data

Method	Univariate imputation						Multivariate imputation					
	Quantitative			Qualitative			Quantitative			Qualitative		
	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR
Hot-deck	✓	×	×	×	×	×	×	×	×	×	×	×
Cold-deck	×	×	✓	×	×	×	×	×	×	×	×	×
LOCF/BOCF/NOCB	×	×	✓	×	×	✓	×	×	×	×	×	×
Mean and Median	×	×	✓	×	×	×	×	×	×	×	×	×
Mode	×	×	✓	×	×	✓	×	×	×	×	×	×
MICE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
KNN	✓	×	✓	✓	×	✓	✓	×	✓	✓	×	✓
Linear regression	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Logistic regression	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MissForest	✓	×	✓	✓	×	✓	✓	×	✓	✓	×	✓
Neural networks	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In the following, we present the main deep learning models used for missing data imputation. These models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), variational autoencoders (VAEs), and GANs. Each model has a unique approach in handling incomplete data.

### 3.3.1. Convolutional neural networks

CNNs [29] are particularly well-suited for imputing missing data in images, where missing pixels can be estimated based on spatial correlations with nearby pixels. CNNs utilize convolutional layers to extract features from input images, effectively capturing local dependencies. This makes them ideal for applications where data exhibits spatial patterns, such as medical imaging or satellite data.

### 3.3.2. Recurrent neural networks

RNNs [30] are frequently employed for imputing temporal data, as they leverage previous information to predict missing values. These models maintain an internal state that captures the sequence of previous inputs, making them suitable for time series imputation. An advanced variant, long short-term memory (LSTM) networks, addresses the vanishing gradient problem by maintaining long-term dependencies, which is particularly useful for long-range temporal correlations.

### 3.3.3. Variational autoencoders

VAEs [31] use an encoder to compress data into a latent representation and a decoder to reconstruct it. Their probabilistic framework enables realistic imputations in complex, non-linear datasets. They achieve this by generating data distributions close to the original.

### 3.3.4. Generative adversarial networks

GANs [32] consist of a generator and a discriminator that compete during training. The generator produces synthetic data, while the discriminator distinguishes real from generated data. This adversarial learning enables realistic imputations for complex data types.

### 3.3.5. Comparative advantages of deep learning models

Deep learning models outperform traditional methods in capturing non-linear and high-dimensional patterns. GANs and VAEs, in particular, generate realistic imputations. However, they require significant computational resources, are sensitive to hyperparameters, and risk overfitting with limited data. Despite these challenges, their feature-learning capability makes them highly effective across various data types.

### 3.3.6. GAN-based models

GANs [33] iteratively improve data generation through competition between a generator and discriminator. This adversarial approach has enabled breakthroughs in missing data imputation [34].

#### a. Generative adversarial imputation network

Generative adversarial imputation network (GAIN) [35] adapts GAN principles for imputation, using a mask matrix to highlight missing values. The generator predicts missing data, while the discriminator evaluates imputations. The architecture involves three components: data, mask, and noise matrices. Algorithm 1 outlines its operation.

---

#### Algorithm 1 Pseudo-code of GAIN

---

**Require:** Dataset with missing values

**Ensure:** Complete data vector

- 1: Initialize generator  $G$  and discriminator  $D$
  - 2: **while** loss has not converged **do**
  - 3:   Draw random samples and masks
  - 4:   Generate imputations with  $G$
  - 5:   Compute discriminator loss and update  $D$
  - 6:   Compute generator loss and update  $G$
  - 7: **end while**
- 

#### b. Missing data GAN

MisGAN [36] learns high-dimensional data distributions by combining two generators and discriminators for masks and data. Algorithm 2 summarizes its training process.

**Algorithm 2** Pseudo-code of MisGAN**Require:** Dataset with missing values**Ensure:** Complete data

- 1: **while** iterations not complete **do**
- 2:   Train mask discriminator  $D_m$  and generator  $G_m$
- 3:   Train data discriminator  $D_x$  and generator  $G_x$
- 4:   Update both generators with combined loss
- 5: **end while**

## c. Other GAN variants

- Stackelberg GAN: uses multiple generators to handle complex imputation tasks [32].
- SolarGAN: tailored for solar data imputation with Wasserstein GAN techniques [37].
- ConvGAIN: extends GAIN with convolutional layers for spatio-temporal correlations [38].
- DEGAIN: builds on GAIN with enhanced loss functions [39].
- GAN-based Sperm-inspired pixel imputation: introduces an identity block and a sperm motility-inspired metaheuristic to improve imputation robustness and address mode collapse and vanishing gradients [40].
- Menstrual cycle inspired GAN : integrates adaptive loss functions and identity blocks inspired by endometrial behavior to enhance imputation in medical images [41].

Deep learning, particularly GANs, provides powerful tools for imputing missing data. Despite challenges like high computational demands and overfitting risks, ongoing innovations continue to improve their robustness and adaptability across various domains.

## 4. EVALUATION METHODS

Evaluation metrics are essential for measuring the quality of missing data imputation in images by quantifying the discrepancy between the original and imputed data. This work focuses on three main metrics: mean squared error (MSE), root mean squared error (RMSE), and Fréchet inception distance (FID).

### 4.1. Mean squared error

MSE measures the average of the squared differences between the actual and imputed values. A lower MSE indicates better imputation quality. A key variant of MSE is RMSE, which computes the square root of the average squared prediction errors:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\text{MSE}(y, \hat{y})} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

RMSE is often preferred for evaluating imputation models as it: Provides error measurements in the same units as the target variable, aiding interpretation, Penalizes larger errors more significantly, and Is less sensitive to outliers compared to MSE.

### 4.2. Fréchet inception distance

The FID, introduced by [10], is widely used to evaluate the quality of images generated by generative models, including GANs. It has been applied to state-of-the-art models such as StyleGAN1 and StyleGAN2 [42]. FID quantifies the similarity between the feature distributions of generated and real images. It calculates the Fréchet distance between two probability distributions. FID provides a robust measure for assessing the fidelity of generative models by comparing how closely generated images match real image distributions.

### 4.3. Evaluation framework

This work employs the following metrics to evaluate missing data imputation quality: i) MSE and RMSE: assess prediction accuracy and variability. ii) FID: evaluates the fidelity of generative models, especially GANs. These metrics establish a strong foundation for selecting and optimizing imputation models in various contexts. The subsequent section will analyze imputation models, highlighting their strengths, limitations, and practical applications.

## 5. DISCUSSION

This section aims to provide a critical evaluation of the methods discussed. This section evaluates the methods for missing data imputation based on three main criteria: the imputation approach (single or multiple), the variable types (quantitative or qualitative), and the missing data mechanisms (MCAR, MAR, or MNAR). Table 3 summarizes these methods, illustrating their applicability and limitations.

- Traditional methods: hot-deck and cold-deck approaches perform well in specific scenarios (MCAR, MAR) but fail in complex mechanisms (MNAR) or when continuous variables are involved. Mean and median imputations are effective under MCAR but introduce bias in MAR and MNAR cases.
- Machine learning approaches, such as KNN and regression exhibit adaptability to both quantitative and qualitative variables. However, their performance declines in MNAR cases or non-linear relationships.
- Advanced models: neural networks and MICE provide the most comprehensive solutions, excelling across all criteria, including the ability to handle diverse data types and multiple imputations.

GAN-based models: Table 4 presents a detailed comparison of GAN-based models, showcasing their architectures, evaluation metrics, and domain-specific applications. Key insights include: i) GAIN: offers a flexible, fully connected architecture effective for categorical, numerical, and image data. Extensions to temporal and textual domains are recommended. ii) view imputation generative adversarial network (VIGAN): focused on image data with multimodal DAE and CNN. Its performance could improve with multi-view datasets. iii) SolarGAN: designed for time-series data, with potential applications in photovoltaic forecasting.

In conclusion, neural networks and GAN-based models stand out for their robustness and adaptability. However, careful alignment of method selection with data type and missing data mechanism is crucial. Future research should emphasize domain-specific optimizations and comparisons to address complex scenarios effectively.

Table 4. Comparison of GAN-based models

Model	Year	Dataset	Evaluation	Code	Internal structure			Missing data	
					Architecture	G	D	Mechanism	Type
GAIN	2018	UCI and MNIST	RMSE	Yes	FC	1	1	MCAR	Qualitative
VIGAN	2017	MNIST	RMSE	Yes	FC, CNN	2	2	NA	Quantitative
MisGAN	2019	CIFAR-10 and CelebA	FID and RMSE	Yes	FC, CNN	2	2	MCAR	Quantitative
CollaGAN	2019	T2-FLAIR and RaFD	NMSE and SSIM	Yes	CNN	1	1	NA	Quantitative
Stackelberg	2018	Tiny ImageNet	FID	No	FC	M	1	NA	Quantitative
SolarGAN	2020	GEFCom2014	MSE	Yes	GRUI, FC	1	1	NA	Qualitative
ConvGAIN	2021	CHS dataset	RMSE	Yes	CNN	1	1	MCAR	Qualitative
DEGAIN	2022	UCI	RMSE and FID	No	Deconv	1	1	NA	NA
GSIP	2025	Energy Images, NREL Solar Images, and NREL Wind Turbine	RMSE, RSNR, SSIM, FID	No	CNN, Deconv	1	1	NA	Qualitative
MCI-GAN	2025	Medical images	RMSE, RSNR, FID, IS, SSIM	NO	CNN	1	1	MAR	Quantitative

Table 4 provides an overview of GAN-based models for missing data imputation. It compares their internal structures, architectures, evaluation metrics, tested datasets, and data handling capabilities across various domains (categorical, numerical, image, and time series). This analysis offers a detailed understanding of each model's strengths, limitations, and application potential.

Key insights: among GAN-based models, GAIN stands out for its flexibility and broad applicability across categorical, numerical, and image data. VIGAN leverages multimodal DAE and CNN for image tasks, with room for multi-view improvements. MisGAN performs well under MCAR but requires adaptation for broader use. CollaGAN focuses on image-to-image translation, while Stackelberg GAN explores multi-generator designs for numerical data. SolarGAN is tailored to time-series imputation, and ConvGAIN and DEGAIN enhance spatial and generator performance through CNNs and deconvolution.

Overall, these models illustrate the evolution of GAN-based imputation. GAIN, in particular, provides a strong base for future domain-specific extensions. Emphasis should be placed on improving adaptability and addressing stability and interpretability challenges.

### 5.1. Best-performing methods by missing data mechanism

Based on the literature synthesis and the comparative Table 3, the following conclusions can be drawn:



- MCAR: simple statistical methods such as mean/median imputation and KNN are often sufficient due to the randomness of missingness. GAN-based models like GAIN and MisGAN also perform well under MCAR assumptions.
- MAR: more advanced methods such as MICE, MissForest, and neural networks are better suited, as they can leverage observed variable relationships. GAN models like MCI-GAN also show promising results.
- MNAR: handling MNAR remains challenging. Methods based on neural networks and certain robust variants of GANs (e.g., DEGAIN, GSIP) offer improved results, though no method fully resolves the MNAR scenario without domain knowledge or additional assumptions.

## 5.2. Challenges and limitations of GAN-based imputation models

Despite their powerful capabilities, GAN-based imputation models face several technical challenges that limit their reliability and generalizability.

### 5.2.1. Mode collapse and convergence issues

GAN training is notoriously unstable due to the adversarial nature of the generator and discriminator. Mode collapse, where the generator produces limited data patterns regardless of input noise, results in biased or unrealistic imputations. Additionally, convergence is difficult to assess, and training may oscillate or diverge without providing meaningful imputations [43].

### 5.2.2. Hyperparameter sensitivity

GANs are sensitive to hyperparameters such as learning rates, batch sizes, and architecture depth. Fine-tuning these parameters is often problem-specific and computationally expensive, requiring extensive empirical experimentation [44]. Poorly chosen hyperparameters may lead to overfitting or non-convergent training, particularly when working with sparse datasets or complex data structures.

### 5.2.3. Potential solutions

Several strategies have been proposed to improve the stability and effectiveness of GAN-based imputation: i) Pretraining techniques: pretraining the generator or discriminator with autoencoder structures or VAEs can stabilize learning and prevent early collapse [45]. ii) Hybrid architectures: models combining GANs with VAEs (e.g., VAE-GAN) or transformer encoders enhance both stability and representational richness [46]. iii) Regularization and loss design: advanced loss functions (e.g., Wasserstein loss with gradient penalty) and spectral normalization can improve convergence and reduce sensitivity to hyperparameters. iv) Méta-apprentissage: using meta-learning to adaptively select the best imputation strategy depending on the missingness mechanism (MCAR, MAR, MNAR) and the data type has shown promise in improving generalizability. These improvements not only enhance imputation quality but also address ethical and interpretability concerns by making GANs more stable, transparent, and adaptable to real-world constraints.

## 5.3. Ethical implications of data imputation

Data imputation techniques, while essential for maintaining data integrity, pose significant ethical challenges, especially when applied in critical domains such as healthcare, finance, and social sciences. The use of advanced imputation methods, particularly those based on GANs, raises concerns related to accuracy, fairness, transparency, and accountability.

### 5.3.1. Risk of inaccurate imputation

One of the primary ethical concerns in data imputation is the risk of inaccurate imputations leading to erroneous conclusions or biased decision-making. In healthcare, for instance, imputing missing patient data with GAN-based methods without adequate validation could result in misleading diagnostic outcomes or inappropriate treatments [47]. In finance, incorrect imputation of financial metrics might lead to flawed credit scoring, adversely affecting individuals or businesses [48].

### 5.3.2. Fairness and bias

GAN-based imputation methods may inadvertently propagate or amplify existing biases present in the training data. For example, if demographic data from underrepresented groups are underimputed or inaccurately generated. It can lead to discriminatory outcomes in automated decision-making systems, such as loan approvals or health risk assessments [49].

### 5.3.3. Opacity and lack of interpretability

GANs are often considered "black-box" models, meaning their decision-making processes are inherently difficult to interpret. This lack of transparency poses ethical challenges when imputations significantly influence high-stakes decisions. Developing interpretable imputation models or integrating explainable AI (XAI) techniques is essential to ensure accountability and build trust in automated systems [50].

### 5.3.4. Privacy concerns

The use of GANs for data imputation may also raise privacy issues. Since GANs generate synthetic data that resemble real-world data, there is a risk that sensitive information might be reconstructed, even when anonymization techniques are applied. This potential for data leakage necessitates rigorous privacy-preserving mechanisms during the imputation process [51].

### 5.3.5. Mitigation strategies

To address these ethical challenges, researchers and practitioners should consider the following approaches: i) ethical guidelines for data imputation: establishing clear guidelines to evaluate the ethical impact of imputation methods, particularly in sensitive domains. ii) Algorithmic fairness audits: regularly auditing GAN-based models to identify and mitigate bias, especially when handling demographic data. iii) Improving model transparency: incorporating XAI methods, such as feature attribution and latent space visualization, to make imputed results more interpretable and trustworthy. iv) Data privacy mechanisms: employing techniques like differential privacy to ensure that GAN-generated data does not inadvertently reveal personal information.

## 6. CONCLUSION AND FUTURE WORK

This study underscores the significance of selecting imputation methods that are well-suited to the nature of missing data and variable types. GAN-based models have demonstrated strong potential in handling complex data structures such as images and time series, especially in high-impact fields like healthcare, finance, and environmental analysis. Their adaptability and capacity to generate realistic values make them valuable tools in advancing missing data imputation techniques. However, these models still face notable challenges including training instability, mode collapse, and hyperparameter tuning difficulties. Hybrid models that combine GANs with VAEs have emerged as a promising direction, offering both the generative strength of GANs and the stability of VAEs. Moreover, the integration of meta-learning techniques could allow for dynamic selection of imputation strategies based on dataset characteristics, thus enhancing generalization. Despite their performance, the interpretability of GAN-based models remains limited, raising concerns in critical domains where transparency is essential. Future research should therefore explore the incorporation of XAI methods to improve understanding and trust in the imputation process. Additionally, efforts should focus on scaling these models for real-world applications, improving their computational efficiency, and ensuring their reliability across diverse data contexts. Overall, this work lays the groundwork for further exploration into robust, interpretable, and scalable imputation strategies using GANs.

## ACKNOWLEDGMENTS

The authors would like to sincerely thank Mr. Mohamed El Hadramy Oumar, founder of Vector Mind, for his generous support in facilitating the transaction required for the publication process. His assistance is gratefully acknowledged.

## FUNDING INFORMATION

The authors declare that no funding was involved in the preparation of this manuscript.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rifaa Sadegh	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Ahmed Mohameden	✓	✓		✓		✓	✓			✓	✓			
Mohamed Lemine Salihi	✓			✓		✓		✓	✓	✓		✓		
Mohamedade Farouk Nanne	✓	✓				✓	✓			✓	✓	✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review &amp; Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest related to this work.

## DATA AVAILABILITY




This study did not involve the generation or analysis of new datasets. All referenced data are publicly available and properly cited.

## REFERENCES




- [1] B. K. Beaulieu-Jones and J. H. Moore, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Pacific Symposium on Biocomputing*, World Scientific, 2017, pp. 207–218, doi: 10.1142/9789813207813.0021.
- [2] N. S. Al-Amri and A. M. Subyani, "Analysis of rainfall, missing data, frequency and PMP in Al-Madinah area, western Saudi Arabia," in *Arabian Plate and Surroundings: Geology, Sedimentary Basins and Georesources*, Cham, Switzerland: Springer, 2020, pp. 235–248, doi: 10.1007/978-3-030-21874-4\_9.
- [3] B. Fallah, K. T. W. Ng, H. L. Vu, and F. Torabi, "Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation," *Waste Management*, vol. 116, pp. 66–78, Oct. 2020, doi: 10.1016/j.wasman.2020.07.034.
- [4] J. Ma, J. C. P. Cheng, F. Jiang, W. Chen, M. Wang, and C. Zhai, "A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data," *Energy and Buildings*, vol. 216, June 2020, doi: 10.1016/j.enbuild.2020.109941.
- [5] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.
- [6] R. H. H. Groenwold and O. M. Dekkers, "Missing data: The impact of what is not there," *European Journal of Endocrinology*, vol. 183, no. 4, 2020, pp. E7–E9, doi: 10.1530/EJE-20-0732.
- [7] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5–37, 2010, doi: 10.1016/j.jsp.2009.10.001.
- [8] D. A. Bennett, "How can I deal with missing data in my study?," *Australian and New Zealand Journal of Public Health*, vol. 25, no. 5, pp. 464–469, Oct. 2001, doi: 10.1111/j.1467-842X.2001.tb00294.x.
- [9] D. B. Rubin and N. Schenker, "Multiple imputation in health-care databases: An overview and some applications," *Statistics in Medicine*, vol. 10, no. 4, pp. 585–598, Apr. 1991, doi: 10.1002/sim.4780100410.
- [10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, pp. 6626–6637, 2017.
- [11] T. G. Nick, "Descriptive statistics," *Methods in Molecular Biology*, vol. 404, pp. 33–52, 2007, doi: 10.1007/978-1-59745-530-5\_3.
- [12] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, Apr. 2010, doi: 10.1111/j.1751-5823.2010.00103.x.
- [13] J. S. Haukoos and C. D. Newgard, "Advanced statistics: Missing data in clinical research-part 1: an introduction and conceptual framework," *Academic Emergency Medicine*, vol. 14, no. 7, pp. 662–668, May 2007, doi: 10.1197/j.aem.2006.11.037.
- [14] M. G. Kenward and G. Molenberghs, "Last observation carried forward: A crystal ball?," *Journal of Biopharmaceutical Statistics*, vol. 19, no. 5, pp. 872–888, Aug. 2009, doi: 10.1080/10543400903105406.
- [15] J. Shao, D. C. Jordan, and Y. L. Pritchett, "Baseline observation carry forward: reasoning, properties, and practical issues," *Journal of Biopharmaceutical Statistics*, vol. 19, no. 4, pp. 672–684, Jun. 2009, doi: 10.1080/10543400902964118.
- [16] K. Unnebrink and J. Windeler, "Sensitivity analysis by worst and best case assessment: Is it really sensitive?," *Therapeutic Innovation & Regulatory Science*, vol. 33, no. 3, pp. 835–839, Jul. 1999, doi: 10.1177/009286159903300324.
- [17] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: A comparison of methods," *Journal of Clinical Epidemiology*, vol. 56, no. 10, pp. 968–976, Oct. 2003, doi: 10.1016/S0895-4356(03)00170-7.
- [18] M. Hajja, "Some elementary aspects of means," *International Journal of Mathematics and Mathematical Sciences*, vol. 2013, 2013, doi: 10.1155/2013/689560.
- [19] E. W. Weisstein, "Statistical median," *MathWorld*, 2016, [Online]. Available: <http://mathworld.wolfram.com/StatisticalMedian.html>
- [20] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Applied Artificial Intelligence*, vol. 32, no. 2, pp. 186–196, 2018, doi: 10.1080/08839514.2018.1448143.
- [21] M. K. Faradj, "Which mean do you mean? An exposition on means," *M.Sc. thesis*, Department of Mathematics, Louisiana State University, Baton Rouge, United States, 2004, doi: 10.31390/gradschool.theses.1852.

- [22] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993, doi: 10.1080/01621459.1993.10476408.
- [23] S. V. Buuren and C. G. M. Oudshoorn, *Multivariate imputation by chained equations: MICE V1.0 User's manual*, TNO Report, Leiden, Netherlands, 2000.
- [24] S. V. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.
- [25] J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105–115, Oct. 2010, doi: 10.1016/j.artmed.2010.05.002.
- [26] L. Zhao, Y. Chen, and D. W. Schaffner, "Comparison of logistic regression and linear regression in modeling percentage data," *Applied and Environmental Microbiology*, vol. 67, no. 5, pp. 2129–2135, May 2001, doi: 10.1128/AEM.67.5.2129-2135.2001.
- [27] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Medical Informatics and Decision Making*, vol. 16, no. S3, Jul. 2016, doi: 10.1186/s12911-016-0318-z.
- [28] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Oct. 2012, doi: 10.1093/bioinformatics/btr597.
- [29] Y. Zhuang, R. Ke, and Y. Wang, "Innovative method for traffic data imputation based on convolutional neural network," *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 605–613, Nov. 2019, doi: 10.1049/iet-its.2018.5114.
- [30] M. Sangeetha and M. Senthil Kumaran, "Deep learning-based data imputation on time-variant data using recurrent neural network," *Soft Computing*, vol. 24, no. 17, pp. 13369–13380, 2020, doi: 10.1007/s00500-020-04755-5.
- [31] Q. Ma, X. Li, M. Bai, X. Wang, B. Ning, and G. Li, "MIVAE: Multiple imputation based on variational auto-encoder," *Engineering Applications of Artificial Intelligence*, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106270.
- [32] H. Zhang and D. P. Woodruff, "Medical missing data imputation by stackelberg GAN," *Machine Learning Department, DAP Report*, Carnegie Mellon University, pp. 1–13, 2018.
- [33] I. Goodfellow et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.
- [34] J. Kim, D. Tae, and J. Seok, "A survey of missing data imputation using generative adversarial networks," in *2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020*, IEEE, Feb. 2020, pp. 454–456, doi: 10.1109/ICAIIIC48513.2020.9065044.
- [35] J. Yoon, J. Jordon, and M. Van Der Schaar, "Supplementary materials GAIN: Missing data imputation using generative adversarial nets," in *35th International Conference on Machine Learning, ICML 2018*, PMLR, 2018, pp. 9052–9059.
- [36] S. C. X. Li, B. Jiang, and B. M. Marlin, "MisGAN: Learning from incomplete data with generative adversarial networks," *7th International Conference on Learning Representations, ICLR 2019*, 2019, pp. 1–20.
- [37] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, "SolarGAN: Multivariate solar data imputation using generative adversarial network," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 743–746, Jan. 2021, doi: 10.1109/TSTE.2020.3004751.
- [38] E. Adeli, J. Zhang, and A. A. Taflanidis, "Convolutional generative adversarial imputation networks for spatio-temporal missing data in storm surge simulations," *arXiv-Computer Science*, pp. 1–32, 2021.
- [39] R. Shahbazian and I. Trubitsyna, "DEGAN: Generative-adversarial-network-based missing data imputation," *Information*, vol. 13, no. 12, Dec. 2022, doi: 10.3390/info13120575.
- [40] G. M. Mahmoud, W. Said, M. M. Fadel, and M. Elbaz, "Novel GSIP: GAN-based sperm-inspired pixel imputation for robust energy image reconstruction," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-82242-9.
- [41] H. S. Marie and M. Elbaz, "MCI-GAN: a novel GAN with identity blocks inspired by menstrual cycle behavior for missing pixel imputation," *Neural Computing and Applications*, vol. 37, no. 16, pp. 9669–9703, Mar. 2025, doi: 10.1007/s00521-025-11059-y.
- [42] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2020, pp. 8107–8116, doi: 10.1109/CVPR42600.2020.00813.
- [43] M. Megahed and A. Mohammed, "Multi-GANs with shared generator: an approach for handling mode collapse issue," in *6th International Conference on Computing and Informatics, ICCI 2024*, IEEE, Mar. 2024, pp. 483–489, doi: 10.1109/ICCI61671.2024.10485012.
- [44] T. S. Rodrigues and P. R. Pinheiro, "Hyperparameter optimization in generative adversarial networks (GANs) using gaussian AHP," *IEEE Access*, vol. 13, pp. 770–788, 2025, doi: 10.1109/ACCESS.2024.3518979.
- [45] M. Jaweed and R. F. Shaikh, "Optimizing generative AI by overcoming stability mode collapse and quality challenges in GANs and VAEs," *MSW Management Journal*, vol. 34, no. 2, pp. 497–507, 2024.
- [46] R. Gupta, S. Tiwari, and P. Chaudhary, "Large generative models for different data types," in *Generative AI: Techniques, Models and Applications*, Springer Nature Switzerland, 2025, pp. 103–162, doi: 10.1007/978-3-031-82062-5\_6.
- [47] S. Nayak and P. M. Khilar, "Data imputation in healthcare applications," in *AI Healthcare Applications and Security, Ethical, and Legal Considerations*, IGI Global, 2024, pp. 49–67, doi: 10.4018/979-8-3693-7452-8.ch004.
- [48] D. S. Nkambule, B. Twala, and J. H. C. Pretorius, "Effective machine learning techniques for dealing with poor credit data," *Risks*, vol. 12, no. 11, Oct. 2024, doi: 10.3390/risks12110172.
- [49] K. Fujisawa, R. Takeda, and A. Mori, "AI for credit risk modeling: a deep learning approach," *Multidisciplinary Journal of Engineering and Technology*, vol. 2, no. 1, 2025, doi: 10.61784/mjet3023.
- [50] E. ŞAHİN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Computing and Applications*, vol. 37, no. 2, pp. 859–965, Nov. 2025, doi: 10.1007/s00521-024-10437-2.
- [51] A. Mishra, A. Majumder, D. Kommineni, C. Anna Joseph, T. Chowdhury, and S. K. Anumula, "Role of generative artificial intelligence in personalized medicine: a systematic review," *Cureus*, Apr. 2025, doi: 10.7759/cureus.82310.




**BIOGRAPHIES OF AUTHORS**

**Rifaa Sadegh**    received her M.S. degree in Computer Science from the University of Nouakchott in 2019 and her Ph.D. in the same field from the same university in 2023. She is currently a researcher and developer specializing in artificial intelligence and software engineering. Her current research interests include deep learning, AI-driven solutions for education, and public sector innovation. She can be contacted at email: rifasadegh@gmail.com.






**Ahmed Mohameden**    received his M.S degree in Distributed Information Systems and Ph.D. in Data Mining from Cheikh Anta Diop University in 2014 and 2018, respectively. He is currently an Assistant Professor at University of Nouakchott and IT bachelor coordinator. His current research interests include deep learning, generative networks, dynamic ego-community detection, and complex networks structures. He can be contacted at email: amed.mohameden@gmail.com.



**Mohamed Lemine Salihi**    had his Ph.D. in Numerical Simulations, Applied Mathematics from the Joseph Fourier University Grenoble, France in 1998. Since 2006 he is a teacher researcher at the University of Nouakchott in Mauritania. His current research interests include deep learning, blockchain, and big data. He can be contacted at email: mlsalihi@gmail.com.



**Mohamedade Farouk Nanne**    is currently a Full Professor at University of Nouakchott, Mauritania. He received a Ph.D. in Computer Science from Paris 8 University Vincennes-Saint-Denis, France. In 2012, he joined the University of Nouakchott as an Assistant Professor. He is habilitated to supervise research (HDR) since 2016. His research mainly focuses on deep learning, E-learning, blockchain, and big data. He can be contacted at email: mohamedade@gmail.com.