

Deep Generative Models for Missing Data Imputation under MCAR, MAR, and MNAR Mechanisms: A Comprehensive Survey

Hiba Z. Mohammed¹, Wael Hadeed²

¹Hiba Z. Mohammed, Master's student, Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

²Wael Hadeed, Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

¹hiba.25csp78@student.uomosul.edu.iq · ²wael.hadeed@uomosul.edu.iq

Received: August 25,2026 Revised: October 04,2026 Accepted October 11,2026

Abstract:

Missing-data imputation has become a necessary task to address as missing data problems have become prevalent in modern datasets, and their presence can be detrimental to statistical inference, machine-learning algorithms, and predictive performance. State-of-the-art imputation approaches often lack flexibility in the modeling of nonlinear relationships and dependencies. This provides motivation for deep generative approaches that can first learn the underlying data distribution to generate candidate imputations for missing observations. This survey covers recent advances in deep generative missing-data imputation approaches, including autoencoders, GANs, diffusion approaches, and related attention, graph, and hybrid models. We provide an overview of studies on multiple datasets, including healthcare, time-series, spatiotemporal, educational, and genomic datasets, and compare methodologies based on modeling rationale, merits and setbacks, and downstream prediction implications. The results demonstrate that diffusion based approaches are promising for temporal and spatiotemporal contexts, VAEs offer flexible probabilistic representations, and GANs excel at distribution matching, but computational efficiency, accounting for uncertainty, MNAR mechanisms, scalability, privacy preservation, and predictive performance limitations persist. The study highlights directions for future methodology, including prediction-centric imputation, scalable models, and privacy protection.

Keywords: *Missing Data, Data Imputation, Deep Learning, Generative Models, Variational Autoencoder, Generative Adversarial Network, Diffusion Models, Time Series, Prediction.*

1 Introduction

Datasets often contain missing data which is a big issue in the fields like healthcare, time-series analysis, education, energy, and genetics. If some information is missing it could be caused by various reasons for example, if the data has simply not been collected properly or if there has been a mistake in measurements, a sensor has failed, it has happened while transmitting data, or the nature of the way data was collected itself.

Imperceptibility of the data may significantly decrease the level of trustworthiness of the statistical analysis and machine learning models, besides it may even result in poor performance of prediction downstream. Hence, having

good and reliable missing data filling methods is important as they preserve data quality and ensure that statistical and modeling works can be reliable and effective.

Among the traditional imputation methods used frequently along with their simplicity are statistical and regression-based ones. Nonetheless, they may have a limited capacity to uncover the nonlinear features and the complex dependencies in high-dimensional or complex-simulated data. The new era in machine learning brought about deep learning which led to the development of generative methods such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models that can model complex data and provide realistic values for the missing data.

Corresponding author: Hiba Z. Mohammed: hiba.25csp78@student.uomosul.edu.iq

The issue of dependable data processing is highlighted even more so by application fields, for example cybersecurity, IoT, machine learning-based prediction ((Albasheer and Abdullah, 2025);(Qasha and Sulyaman, 2025)). Likewise, the application-based deep learning reconstruction techniques have been developed by medical scientists to retrieve meaningful features useful for the analysis and prediction after the classification of the medical data (Najeeb and Dahl, 2022). These application domains point out the necessity of having trustworthy representations of the data and recovery when missingness influences subsequent analytical jobs.

This study specifically examines deep generative methods that deal with missing- data imputation in missing completely at random (MCAR) missing at random (MAR) and missing not at random (MNAR). Mechanisms It discusses generative autoencoder (VAEs) generative adversarial networks (GANs) and diffusion processes, attention graph and hybrid Models along with their characteristics for imposing variability between missing pattern and imputation methods and data type. Specific points such as sources of variability in imputation differences between methods and downstream impacts are being highlighted.

Research Question

One research question this paper seeks to answer is: “How do deep generative models for missing-data imputation perform under MCAR, MAR, and MNAR mechanisms across different data types and application domains, and what are their main advantages, limitations, and future research directions?”

Objectives of the Survey

The main goals of this survey are to:

1. Summarize the literature of the recent deep generative approaches for missing-data imputation, including VAEs, GANs and diffusion-based models;
2. Assess if they are effective under MCAR, MAR, and MNAR mechanisms;
3. Analyze how the quantitative performance aspects of their results compare with different datasets and application areas.
4. Pinpoint their key weaknesses and possible development avenues.

2 REVIEW METHODOLOGY

This survey used a structured literature-review approach to identify and analyze studies on deep generative models for missing-data imputation under MCAR, MAR, and MNAR mechanisms. Recent research was prioritized, while seminal and relevant review or benchmarking studies were retained when necessary.

Literature Search: Searches were conducted using Google Scholar, Scopus, Web of Science, Elsevier/ScienceDirect, and ResearchGate as a supplementary source. Keywords included *missing data*, *missing-data imputation*, *deep generative models*, *VAE*, *GAN*, *GAIN*, *diffusion models*, *MCAR*, *MAR*, and *MNAR*.

Study Selection: Studies were included when they addressed missing-data imputation using deep learning or generative approaches and provided relevant methodological or evaluation information. Studies outside the scope of missing-data imputation, duplicates, and publications with insufficient methodological information were excluded.

Following screening, **25 studies** were selected and classified according to missingness mechanism, data type, imputation approach, main model, and evaluation metrics.

3 Missing Data

Missing data means there are gaps or voids in the database when a value that should have been recorded is missing or not captured at a particular point in temporal, spatial, and multimodal data. Real-world situations are such that data loss can happen due to various causes like malfunctions of sensors and data not being collected completely. Transmission problems or characteristics that are inherent to the process of data generation could also be reasons (Afkanpour et al., 2024), (Kazijevs and Samad, 2023), (Buwaneswaran, 2024).

Missing data effects not only change with the number but also with the type of the mechanisms that cause the absence. As a result, to come up with a suitable imputation technique we must have a profound knowledge of the missingness mechanism.

3.1 Missing Completely at Random (MCAR)

In a Missing Completely at Random (MCAR) condition, the likelihood of a data element being missing is not dependent on what other data elements have been or have not been observed. In other words, missingness occurs randomly and is not influenced by the characteristics of the observations or by other variables. For example, if some measurements are randomly lost due to an unexpected technical error or random data-recording failure, the missing values can be considered MCAR. This mechanism represents the simplest form of missingness because the observed data remain statistically representative of the complete dataset (Buwaneswaran, 2024), (Chen *et al.*, 2023), (Chen *et al.*, 2024).

3.2 Missing at Random (MAR)

Under the Missing at Random (MAR) mechanism, the probability that a value is missing may depend on other information that is observed in the dataset, but not directly on the missing value itself after accounting for the observed variables. Thus, the available variables can provide useful information about the missing observations and help estimate their likely values. For example, if a patient's laboratory measurement is missing and the probability of missingness is related to the patient's age or previously record clinical measurements, the data may follow an MAR mechanism. Compared with MCAR, MAR represents a more realistic and complex missing-data scenario because the observed variables contain information about the missingness process. This makes the appropriate use of observed features particularly important for accurate imputation (Alcaraz *et al.*, 2023), (Kazijevs and Samad, 2023).

3.3 Missing Not at Random (MNAR)

The Missing Not at Random (MNAR) mechanism represents a more challenging missing-data scenario in which the probability of a value being missing is directly related to the unobserved value itself or to information that remains unavailable. On its own, observed data might not carry enough clue to precisely explain patterns of absent observations. For this reason only reliance on observed variables to perform imputations via usual methods can lead to biased or inaccurate estimate of the missingness. It is quite complicated to get rid of a missingness pattern when such pattern is MNAR; it requires one at the same time taking into account a missingness pattern and a data-generation pattern. At least some types of statistical models are needed to identify the

relationships within the partial dataset while the factors that led to missing data can also be identified separately.

(Chen *et al.*, 2024), (Buwaneswaran, 2024) have put forward a deep generative method of imputing missing data for MNAR. They modelled the incomplete variable and the missingness mechanism within an integrated deep generative process. Similar to them (Pereira *et al.*, 2024) looked into imputation of missing values under a MNAR setup with the help of a multi-method benchmarking analysis. There is evidence to show that imputation techniques vary greatly depending on the nature and level of missingness present in the data. These works are a clear example of the fact that one has to fully acknowledge and factor in the missingness generation process while building powerful and robust imputation methods for very complex real world datasets.

3.4 Classification of Missing Data Imputation Methods

3.4.1 Traditional Imputation Methods

Traditional missing data handling imputation strategies still are useful as simple and less time-consuming approaches for dealing with missing data. In the reviewed studies, they were mainly used as baseline methods for evaluating advanced machine-learning and generative approaches. The most frequently reported methods include mean, median, K-Nearest Neighbors (KNN), and multiple imputation.

3.4.1.1 Mean and Median Imputation

Mean imputation replaces a missing numerical value with the mean of the observed values, whereas median imputation uses the middle value of the observed distribution. These methods are simple and computationally inexpensive and can provide useful baseline results when the amount of missing data is relatively limited. However, they do not capture relationships between variables and may reduce the variability of the completed data. In the reviewed studies, mean and median methods were mainly used as comparison baselines for more advanced imputation approaches (Kazijevs and Samad, 2023), (Kowsar, 2025), (Wang *et al.*, 2023b), (Xu *et al.*, 2026).

3.4.1.2 Regression-Based Imputation

Regression-based imputation estimates missing values by exploiting relationships between the incomplete variable and other observed variables. The model is trained using available observations and then used to predict the missing components. This approach can be more informative than simple statistical replacement when meaningful relationships exist among variables. However, conventional regression models may have limited ability to represent complex nonlinear dependencies, which motivates the use of more flexible machine-learning and deep-learning approaches.

3.4.1.3 K-Nearest Neighbors (KNN) Imputation

KNN imputation estimates a missing value using information from the most similar observations. The method identifies neighboring samples according to their observed features and uses their corresponding values to estimate the missing component. KNN can capture local relationships within a dataset and is therefore more flexible than simple mean or median replacement. However, its computational cost may increase with dataset size, and its effectiveness depends on the distance measure and number of neighbours. KNN was used as a comparison method in several of the reviewed studies, particularly in healthcare and structured datasets (Afkanpour *et al.*, 2024), (Kuang *et al.*, 2025), (Liu *et al.*, 2023), (Pereira *et al.*, 2024), (Wang *et al.*, 2023a), (Xu *et al.*, 2026), (Zhang, 2024).

3.4.1.4 Multiple Imputation

Multiple imputation generates several plausible values for missing observations and produces multiple completed datasets rather than a single imputed dataset. The resulting datasets are then analyzed and combined to account for uncertainty associated with missing values.

This strategy is the superior method to capture overall uncertainty as opposed to single-value imputation techniques, and is still widely used statistical method for the analysis of incomplete data. In the studies reviewed multiple imputation was used as a benchmark or reference mainly in the analysis of clinical or MNAR data sets. (Afkanpour *et al.*, 2024), (Buwaneswaran, 2024), (Pereira *et al.*, 2024).

In general, through these methods, we were able to derive useful as well as interpretable baselines, especially for relatively simpler datasets, with lower rates of missing data. Nevertheless, the capability of the methods to detect nonlinear relationships, temporal dependencies, and complicated data

distributions is very limited; therefore, it's an important driving force for deep-learning and generative approaches.

3.4.2 Deep Generative Models for Missing Data Imputation

Sometimes called 'the art of imitation', deep generative models learn data distributions through training examples and then, they use those learned representations to generate plausible data points from the distribution that they learned. With the advancement of Deep Learning, generative methods have become a viable competitor against traditional statistical modelling techniques. The reason is their ability to not only handle the complex patterns in data and nonlinear relationships, but also to exploit the dependencies in time and space.

Besides, since they can naturally represent the joint probability of a huge number of features or variables, Deep Generative Models are ideally suited for the analysis of very large-scale temporal as well as heterogeneous datasets. There are several Deep Generative Models that one can use for missing data imputation and one of the main approaches high. In the following section, you will learn about each of them in depth and discover the pros and cons. In addition, a comparison of their performances will be made by a case study in which these types of models will be applied to missing data prediction in real world data. Lighted in the reviewed papers are: VAEs, GANs, and diffusion models.

3.4.2.1 Variational Autoencoder (VAE)-Based Methods

Variational Autoencoders (VAEs) are composed primarily of an encoder and a decoder. The encoder converts observed data into a probabilistic latent representation, and the decoder uses that representation to regenerate the data. During imputation, the learned latent representation provides information that can be used to reconstruct plausible values for missing components. This makes VAE-based methods particularly useful when the relationships among variables are complex and cannot be adequately represented by simple statistical methods.

The basic VAE objective can be represented by Equation (1).

$$\begin{aligned} \mathcal{L}_{\text{VAE}}(\theta; \phi; x) = & \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x \\ & | z)] \\ & - D_{\text{KL}}(q_{\phi}(z | x) \\ & \| p(z)) \end{aligned} \quad (1)$$

where x represents the observed data, z is the latent representation, $q_{\phi}(z|x)$ is the encoder distribution, $p_{\theta}(x|z)$ is the decoder distribution, and DKL measures the difference between the learned latent distribution and the prior distribution.

(Mattei and Freiisen, 2019) proposed MIWAE for imputing incomplete datasets, while other reviewed studies extended VAE-based approaches to genomic, educational, and structured data (Ghosh *et al.*, 2026), (Li, 2024), (Matrood, 2026). VAE-based methods provide probabilistic representations and can capture complex dependencies, although their performance depends on the quality of the learned latent representation and the characteristics of the missingness mechanism.

3.4.2.2 Generative Adversarial Network (GAN)-Based Methods

Generative Adversarial Networks (GANs) use two competing components: a generator and a discriminator. The generator attempts to produce realistic values for missing components, whereas the discriminator tries to distinguish generated values from real observations. Through this adversarial training process, the generator gradually learns to produce more realistic imputations.

the general GAN objective can be represented by Equation (2).

$$\begin{aligned} \min_G \max_D V(D; G) = & \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] \\ & + \mathbb{E}_{z \sim p_z(z)} [\log(1 \\ & - D(G(z)))] \end{aligned} \quad (2)$$

where G represents the generator, D

Represents the discriminator, x represents real observations, and z represents the input noise or latent representation.

for missing-data imputation, this adversarial framework is adapted so that the generator focuses on producing plausible estimates for the missing components (Ghanavi *et al.*, 2021). proposed a class-

aware GAN-based approach for missing network-traffic data, while other reviewed studies applied GAN-based approaches to healthcare, educational, and structured datasets (Kuang *et al.*, 2025), (Liu *et al.*, 2023), (Yoon *et al.*, 2018). GAN-based methods can model nonlinear dependencies effectively, but their training may be sensitive to hyperparameters and can be unstable.

3.4.2.3 Diffusion-Based Methods

Diffusion models use a gradual noise addition and denoising process. During training, noise is progressively added to the data, and the model learns to reverse this process. During imputation, the learned denoising process is conditioned on the observed values to reconstruct plausible missing components.

A simplified forward diffusion process can be represented by Equation (3).

$$\begin{aligned} q(x_t | x_{t-1}) \\ = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}; \beta_t \mathbf{I}) \end{aligned} \quad (3)$$

where here x_t represents the data at diffusion step t , β_t controls the amount of noise added at each step, and $\mathbf{I}$ is the identity matrix

On the flip side, the process of recovering the original data from the noisy representation is attempted in reverse.

the study by (Alcaraz *et al.*, 2023) presented the SSSD a diffusion-based model to impute and forecast time-series using structured state space.

PriSTI is a conditional diffusion framework introduced by (Liu *et al.*, 2023) for the purpose of spatiotemporal data imputation

Observed-value-consistent Diffusion Model is another type of a method to replace missing multivariate time series values introduced by (Wang *et al.*, 2023a) such that the consistency of the generated value with the observed values is maintained. MNAR imputation and Wasserstein-based optimization (Chen *et al.*, 2023); (Chen *et al.*, 2024) etc. the works under the umbrella of other reviewed studies mainly focused on diffusion of numerical tabular and structured information with different data.

Table 1: Comparison of Deep Generative Imputation Model Families

Model family	Basic principle	Main strength	Main limitation	Typical applications in the reviewed studies
VAE-based models	Learn a probabilistic latent representation and reconstruct missing components from the learned representation.	Flexible probabilistic modeling and ability to capture complex dependencies.	Performance depends on the quality of the learned latent representation and model design.	Incomplete datasets, genomic and educational data, time-series, and structured data.
GAN-based models	A generator produces plausible missing values while a discriminator distinguishes generated values from observed data.	Effective modeling of nonlinear relationships and realistic data distributions.	Training can be unstable and sensitive to hyper parameters.	Healthcare, educational data, network traffic, and general missing-data imputation.
Diffusion-based models	Gradually denoise corrupted or incomplete data while conditioning on observed values.	Strong distribution modeling and effective handling of temporal dependencies and uncertainty.	Iterative generation can require considerable computational resources.	Multivariate time-series, spatiotemporal data, smart-meter data, and other structured datasets.
Hybrid / Attention / Graph-based models	Combine generative models with attention mechanisms, graph structures, or other learning components.	Can capture feature, temporal, and structural relationships more effectively.	Greater architectural complexity and computational requirements.	Networked time-series, multivariate data, healthcare, and other structured datasets.

for learning the dependencies between complex variables but it also makes the data generation possible which in this paper is the filling of gaps (imputation). Still the drawback lies in the fact that to generate a new sequence, the model first needs to denoise it iteratively, which means more computing power is needed.

3.4.2.4 Hybrid, Attention-, and Graph-Based Methods

Recent studies have enhanced generative imputation models by incorporating attention mechanisms, graph structures, and hybrid architectures. Attention mechanisms help models focus on informative features or time steps, whereas graph-based approaches represent relationships between connected variables or observations. Hybrid models combine two or more approaches to exploit their complementary strengths.

A simplified attention mechanism can be represented by Equation (4).

$$\text{Attention}(Q; K; V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k represents the dimensionality of the key vectors

Graph-based approaches can further represent relationships among observations or variables through a graph structure, allowing information to propagate between connected components. In the reviewed studies, graph-based and attention-enhanced approaches were applied to complex temporal and networked data, while hybrid models combined generative architectures with additional learning mechanisms (Kazijevs and Samad, 2023), (Liu *et al.*, 2023), (Tashiro *et al.*, 2021). These approaches can provide richer representations of complex dependencies, although they may increase architectural complexity and computational requirements.

4 Evaluation Metrics for Missing Data Imputation

4.1 Imputation Accuracy Metrics:

The performance of missing-data imputation methods is commonly assessed by comparing the imputed values with their corresponding true values. The

reviewed studies use several quantitative metrics to evaluate reconstruction accuracy, depending on the data type and objective of the imputation task. The most frequently reported measures include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Relative Error (MRE), Normalized Root Mean Square Error (NRMSE), and Mean Absolute Percentage Error (MAPE). Some studies also employ application-specific measures or evaluate the effect of imputation on downstream prediction tasks.

- Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average absolute difference between the true and imputed values, as represented by Equation (5).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

where y_i represents the true value and $\hat{y}_i$ represents the imputed value. Lower MAE indicates better imputation accuracy. MAE is easy to interpret and is less sensitive to large errors than squared-error measures

- Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared difference between the true and imputed values as represented by Equation (6).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Because the errors are squared, RMSE gives greater importance to large imputation errors. Therefore, it is useful when large deviations from the true values are particularly undesirable.

- Mean Relative Error (MRE)

Mean Relative Error (MRE) is a metric to measure the average of the relative differences between the true and the imputed values. In other words, it shows the imputation error relative to the size of the true values and is mathematically given by:

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i| + \epsilon} \quad (7)$$

where y_i represents the true value, $\hat{y}_i$ represents the imputed value, and ϵ is a small positive constant used to avoid division by zero. Lower MRE indicates better relative imputation accuracy. Because MRE depends on the magnitude of the true values, it can become unstable when y_i is zero or very close to zero.

Normalized Root Mean Square Error (NRMSE)

NRMSE normalizes RMSE according to the scale or variability of the data, as represented by equation (8).

$$NRMSE = \frac{RMSE}{S} \quad (8)$$

where S represents an appropriate scale measure, such as the standard deviation or range of the observed values. Normalization makes it easier to compare imputation performance across variables or datasets with different scales.

- Mean Absolute Percentage Error (MAPE)

MAPE absolute prediction or imputation error as a percentage of the true value data can be represented by equation (9).

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

Lower MAPE indicates better performance. However, similar to MRE, MAPE can become unstable when the true value is zero or close to zero.

4.2 Downstream Prediction and Classification Metrics

When missing data are imputed as a preprocessing step for a subsequent prediction or classification task, the effectiveness of the imputation method can also be assessed through the performance of the downstream model. In this case, classification metrics like Accuracy, Precision, Recall, and F1-score are

applied to see if the missing data values are filled in a way that the predictive decisions made based on the data are good enough. These metrics, from the practical point of view, show another side of imputation that goes beyond reconstructing missing parts by showing how filling in the missing values affects the final prediction task.

4.2.1 Accuracy

Accuracy is a metric which is the ratio of correct predictions to total predicted data by formula (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy is appropriate when the classes are relatively balanced. However, it may provide a misleading impression when the dataset is highly imbalanced.

4.2.2 Precision

Precision measures the proportion of correctly predicted positive cases among all cases predicted as positive, as represented by Equation (11).

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

A high precision indicates that the model produces relatively few false-positive predictions. This metric is particularly useful when false-positive decisions are costly.

4.2.3 Recall

Recall, also known as sensitivity, measures the proportion of actual positive cases correctly identified by the model, as represented by Equation (12).

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

high recall indicates that the model successfully identifies most of the positive cases and produces relatively few false negatives.

4.2.4 F1-Score

F1-score combines Precision and Recall into a single measure, as represented by Equation (13).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

F1-score is particularly useful for imbalanced classification problems because it considers both false-positive and false-negative errors.

4.2.5 Downstream Evaluation

Some reviewed studies evaluate imputation indirectly by examining whether the completed dataset improves a subsequent machine-learning task. In such cases, Accuracy, Precision, Recall, F1-score, AUC, or other task-specific metrics can be reported. This viewpoint is crucial because a method may generate imputations that are numerically accurate without necessarily improving the ultimate forecasting task.

4.3 Probabilistic and Application-Specific Metrics

Incorporating imputation techniques, some research papers rely on different sets of probabilistic or field-

Related metrics to estimate factors which conventional methods like reconstruction or classifying cannot reveal completely. Log-likelihood measures how accurately a probabilistic model describes the original data distribution, and Wasserstein distance (WASS) quantifies the difference in distributions of original vs. imputed data. On the other hand, some research is based on application-specific measures like biological relevance mainly in the case of genomic data or computational runtime in cases where being efficient takes precedence. Together, these metrics give us a clearer picture of how reliable, realistic and ultimately useful the imputed data is.

5 Classification and Evaluation of the Reviewed Studies

The studies that have been reviewed have been grouped on the basis of missing-data mechanism, data type, imputation method, main model, and the evaluation metrics. The grouping enables the understanding of how the data properties relate to the choice of methods for missing data recovery. Furthermore, it makes it possible for the reviewed

methods to be contrasted not only based on the data types but also in terms of application areas such as

Healthcare, temporal data, spatio-temporal data, genomic data, education data, structured data, etc. The 25 research papers (studies) reviewed are characterized in Table 2.

Table 2: Classification and Evaluation of the Reviewed Studies.

Reference	Missingness Mechanism	Data Type / Application	Imputation Category	Main Method / Model	Evaluation Metrics	Reported Quantitative Result
(Afkanpour et al., 2024)	—	Clinical / Structured	Traditional & ML	Comparative review	—	58 studies analysed; 45% conventional, 31% ML/DL, 24% hybrid (PubMed Central (PMC))
(Alcaraz et al., 2023)	—	Time-Series	Generative	Diffusion + SSM	MAE, RMSE	RMSE = 0.0119 on PTB-XL ECG
(Buwaneswaran, 2024)	—	Smart Meter / Time-Series	Generative	Conditional Diffusion	—	—
(Chen et al., 2023)	MNAR	Tabular	Generative	Deep Generative Model	Reconstruction accuracy	11.65% average improvement in reconstruction accuracy over compared MNAR models
(Chen et al., 2024)	MCAR, MAR, MNAR	General / Structured	Generative	Diffusion	MAE, WASS	Best performance reported at 50% missingness; lower Wasserstein distance
(Ghanavi et al., 2021)	MCAR, MAR, MNAR	Network / Structured	Generative	Deep Generative Model	—	GACN outperformed GAIN; at 20% missingness, accuracy was 90.22% vs. 85.75%.
(Ghosh et al., 2026)	MCAR	Genomic	Generative / Federated	VAE / DAE	Biological relevance	94.73% biological structure preserved vs. 0% for mean imputation
(Hallaji et al., 2021)	MAR	Healthcare	GAN / Hybrid	Adversarial Learning	AUC, F-measure	Ranked 1st in accuracy/F-measure across 5–25% missingness
(Kazijevs and Samad, 2023)	MCAR, MAR	Healthcare Time-Series	Deep Learning	Deep Imputation	NRMSD	5–10% accuracy advantage for piecewise interpolation on weak-correlation signals

(Kowsar, 2025)	MCAR, MAR, MNAR	Tabular / Clinical	Deep Learning	DeepIFSAC – Feature & Sample Attention	NRMSE	15–20% lower reconstruction error than baseline GANs; evaluated up to 90% missingness
((Kuang et al., 2025)	MCAR	Binary Image	Generative	MMISVAE	Imputation Accuracy	Higher log-likelihood than MIWAE; effective beyond 50% missingness
(Li, 2024)	—	Time-Series	GAN / Hybrid	GAN + Bayesian	—	—
(Liu <i>et al.</i> , 2023)	—	Spatiotemporal	Generative	PriSTI	MAE, MSE	Outperformed CSDI and GRIN; exact numerical value not stated in the initial summary
(Matrood, 2026)	MCAR, MAR, MNAR	General / Clinical	Review	Multiple Imputation + DL	—	(review article; no primary experimental result used)
(Mattei and Freisen, 2019)	MAR	MNIST / Medical	VAE	MIWAE	Log-likelihood	Outperformed missForest in 7 of 10 medical benchmarks
(Pereira <i>et al.</i> , 2024)	MNAR	Medical / Structured	Traditional & Generative	Benchmark Analysis	MAE	MICE MAE = 0.076 at 10% missingness
(Tashiro <i>et al.</i> , 2021)	—	Time-Series	Diffusion	CSDI	—	MAE = 0.217, 0.301, 0.481 for healthcare at 10%, 50%, 90%; MAE = 9.60 for air quality (Scribd)
(Veldkamp et al., 2025)	MCAR, MAR	Educational	VAE / Hybrid	VAE-IRT	MSE, Runtime	98% speedup: 16 s vs. 22 min
(Wang et al., 2023b)	MCAR, MAR	Networked Time-Series	VAE / Graph	Graph-Enhanced VAE	MAE, MSE	MAE = 0.007 on PeMS-BA vs. 192.047 for mean imputation
(Wang et al., 2023a)	MCAR, MAR	Multivariate Time-Series	Diffusion	Observed-Value Consistent Diffusion	MAE, MRE	1.6% improvement in MAE and 2.4% reduction in MRE vs. CSDI
(Xiao <i>et al.</i> , 2025)	MCAR, MAR	Time-Series	Diffusion	Boundary-Enhanced Diffusion	MAE, RMSE	MAE = 0.162 vs. 0.244 for CSDI; $\approx 33\%$ lower error
(Xu <i>et al.</i> , 2026)	Random Missingness	Multivariate Time-Series	VAE / Recurrent	T-LSTM-VAE	SMAPE, RMSE	RMSE = 0.1678; summary reports $\sim 50\%$ better SMAPE than TimeGAN
(Yoon et al., 2018)	—	Structured / Tabular	GAN	GAIN	—	RMSE: Breast 0.0546 $\pm$ 0.0006; Spam 0.0513 $\pm$ 0.0016; Letter 0.1198 $\pm$ 0.0005;

						Credit 0.1858±0.0010; News 0.1441±0.0007 (van der Schaar Lab)
(Zhang <i>et al.</i> , 2025)	—	Educational	GAN	Generative Adversarial Imputation	RMSE	Significant reduction in knowledge-tracing error vs. GAIN; exact RMSE not reported in the initial summary
(Zhang, 2024)	—	Healthcare / Big Data	Hybrid Generative	DAE + GAN	RMSE, MAE	Significant improvement in RMSE and MAE vs. traditional and standalone models; exact values require extraction from the paper

Imputation performance varies significantly across datasets, missingness levels, processes, and evaluation measures, according to the given quantitative results. For instance, GAIN found dataset-specific RMSE values ranging from 0.0513 to 0.1858, whereas CSDI reported MAE values of 0.217, 0.301, and 0.481 for 10%, 50%, and 90% missingness in the healthcare dataset, respectively. Similarly, in the reported experimental context, DSDI decreased MAE for CSDI from 0.244 to 0.162. These findings show the importance of quantitative comparability, but they also show that because experimental circumstances vary, cross-study numerical data should not be taken as directly interchangeable. .

6 Analysis and Discussion

6.1 Analysis of Traditional Imputation Methods

Traditional imputation methods remain important as baseline approaches, particularly for structured and clinical datasets. Their performance is influenced by the data characteristics, missingness mechanism, and missing-data rate rather than by the imputation method alone (Afkanpour *et al.*, 2024) emphasized that no single method is universally optimal and that method selection should consider the structure and missingness pattern of the data. Under MNAR conditions, (Pereira *et al.*, 2024) reported that

multiple imputation remained competitive, with MICE achieving an MAE of **0.076 at 10% missingness**. Similarly (Kazijevs and Samad, 2023) found that performance varied across healthcare time-series datasets and missing-data conditions. These findings indicate that traditional methods remain useful as reference baselines, particularly when the data structure is relatively simple, but their effectiveness can decrease as missingness becomes more complex.

6.2 Analysis of Deep Generative Imputation Methods

The reviewed studies show that deep generative methods provide different advantages depending on the data structure and missingness mechanism. VAE-, GAN-, and diffusion-based models should therefore not be considered universally interchangeable.

6.2.1 VAE-Based Approaches

VAE-based methods learn probabilistic latent representations and can model uncertainty in incomplete data. (Mattei and Freiisen, 2019) demonstrated the effectiveness of MIWAE across multiple datasets, while later studies extended VAE-based approaches to genomic, educational, and multivariate data (Ghosh *et al.*, 2026); (Veldkamp *et al.*, 2025); (Xu *et al.*, 2026). Quantitatively reported a substantial computational improvement, reducing processing time from approximately 22 minutes to 16

seconds. (Xu *et al.*, 2026) reported an RMSE of 0.1678 for T-LSTM-VAE. These results suggest that VAE-based approaches can provide flexible probabilistic representations, although their effectiveness remains dependent on the learned latent representation and the characteristics of the missingness pattern.

6.2.2 GAN-Based Approaches

GAN-based approaches aim to generate imputations that resemble the observed data distribution. Studies across healthcare, time-series, educational, and structured datasets demonstrate their ability to model nonlinear dependencies (Hallaji *et al.*, 2021); (Li, 2024); (Zhang, 2024); (Yoon *et al.*, 2018). For example, GAIN reported RMSE values of **0.0546 for Breast, 0.0513 for Spam, 0.1198 for Letter, 0.1858 for Credit, and 0.1441 for News** under its benchmark experiments (Yoon *et al.*, 2018). However, GAN-based approaches remain sensitive to model configuration and training stability, which can affect reproducibility and performance across different missingness conditions.

6.2.3 Diffusion-Based Approaches

Diffusion models have received increasing attention for time-series and spatiotemporal imputation because they can model complex dependencies through conditional generation (Alcaraz *et al.*, 2023); (Buwaneswaran, 2024); (Chen *et al.*, 2024); (Liu *et al.*, 2023); (Tashiro *et al.*, 2021); (Wang *et al.*, 2023); (Xiao *et al.*, 2025). Reported results provide quantitative evidence for their effectiveness (Alcaraz *et al.*, 2023). reported an RMSE of 0.0119 on the PTB-XL dataset, while (Xiao *et al.*, 2025) reported an MAE of 0.162, compared with 0.244 for CSDI. (Wang *et al.*, 2023) also reported a 1.6% improvement in MAE and a 2.4% reduction in MRE compared with CSDI. These results indicate strong performance in temporal and spatiotemporal settings; however, diffusion models generally require greater computational resources than simpler imputation approaches

6.2.4 Comparative Analysis across Missingness Mechanisms

The reviewed evidence indicates that model performance depends strongly on the missingness mechanism. VAE- and GAN-based approaches have been applied across MCAR and MAR settings, while diffusion and specialized generative models have

increasingly addressed complex temporal and MNAR scenarios (Chen *et al.*, 2023); (Chen *et al.*, 2024); (Kowsar, 2025); (Pereira *et al.*, 2024). For example, evaluated DeepIFSAC under MCAR, MAR, and MNAR conditions and considered missingness rates up to 90%. (Chen *et al.*, 2023) specifically addressed MNAR imputation and reported an average 11.65% improvement in reconstruction accuracy over the compared MNAR approaches.

Overall, these findings do not indicate that one generative architecture is universally superior. Instead, performance varies with the missingness mechanism, missingness rate, data structure, and evaluation metric. Therefore, quantitative results should be interpreted within their experimental context rather than used for direct ranking across different datasets and protocols.

6.3 Broader Applications and Implications

The analysis of the data reveals that accurate handling of missing data is essential to the development of data-rich applications. Table 2 provides a collection of works which employed generative models for completing missing values in different areas including healthcare, time-series and spatiotemporal data as well as genomic, educational, and other structured datasets (Alcaraz *et al.*, 2023);(Liu *et al.*, 2023); (Ghosh *et al.*, 2026); (Veldkamp *et al.*, 2025).The use cases indicate that the usefulness of an imputation model is not restricted to reconstructing missing values, because it may as well affect subsequent prediction/classification performance based on the imputed data (Hallaji *et al.*, 2021); (Ghanavi *et al.*, 2021).In addition, sound data processing is an aspect which remains essential in a number of domains like cybersecurity, IoTC, cloud computing, and medical devices. This has been the focus of research papers published recently (Albasheer and Abdullah, 2025); (Qasha and Sulyaman, 2025); (Abdullah and Abdullah, 2025); (Sultan and Abdullah, 2023).

7 CONCLUSIONS AND RECOMMENDATIONS

This survey highlights the fact that the effectiveness of missing-data imputation is heavily influenced by three factors: data characteristics, type of missingness, and the rate of missing-data. The

conventional imputation techniques continue to serve efficiently for the cases of simple and structured datasets, on the other hand, deep generative models are a better choice when dealing with data that can be complex, nonlinear, and temporal. The studies surveyed in this review show that no unique method always gives the better performance for all cases. VAE-GAN- based models are able to generalize well by offering representations for quite different data. Among diffusion models, those that generate sequences seem especially relevant for imputing the missing values of temporal and spatiotemporal datasets. Nevertheless, several problems have yet to be solved regarding the effective management of missing not at random data, the efficiency of computing operations, the ability to scale up, as well as the assessment of how certain a prediction is. Therefore, future research should focus on efficient and scalable generative models, improved handling of MNAR data, hybrid approaches, and evaluation of the impact of imputation on downstream prediction, rather than relying solely on reconstruction accuracy.

Acknowledgements

The authors are very much thankful to the Computer Science Department/ College of Computer Science and Mathematics/ University of Mosul/ IRAQ for their full support to accomplish this research.

REFERENCES

Abdullah, A.A. and Abdullah, D.B. (2025) "A Unified Framework for Real-Time Intrusion Detection and Mitigation using Apache Spark," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 16(4), pp. 642–662. Available at: <https://doi.org/10.58346/JOWUA.2025.I4.037>.

Afkanpour, M., Hosseinzadeh, E. and Tabesh, H. (2024) "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review," *BMC Medical Research Methodology*, 24(1). Available at: <https://doi.org/10.1186/s12874-024-02310-6>.

Albasheer, K.D. and Abdullah, D.B. (2025) "Fog and edge computing and its role in distributed real-time containers: A survey," *AIP Conference Proceedings*, p. 030009. Available at:

<https://doi.org/10.1063/5.0286172>.

Alcaraz, J.M.L. and Strodthoff, N. (2023) "Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models," *Transactions on Machine Learning Research*, 2023-Febru, pp. 1–36.

Buwaneswaran, M. (2024) "Scholarship @ Western Missing Data Imputation for Smart Meters : Conditional Denoising Diffusion Model & Temporally Chained Equations Keywords."

Chen, J., Xu, Y., Wang, P. and Yang Y. (2023) "Deep Generative Imputation Model for Missing Not At Random Data," *International Conference on Information and Knowledge Management, Proceedings*, pp. 316–325. Available at: <https://doi.org/10.1145/3583780.3614835>.

Chen, Z., Li, H., Wang, F., Zhang, O., Xu, H., Jiang, X., Song, Z. and Wang, E. (2024) "Rethinking the Diffusion Models for Numerical Tabular Data Imputation from the Perspective of Wasserstein Gradient Flow," (NeurIPS), pp. 1–54. Available at: <http://arxiv.org/abs/2406.15762>.

Ghanavi, R., Liang, B. and Tizghadam, A. (2021) "Generative Adversarial Classification Network with Application to Network Traffic Classification," *Proceedings - IEEE Global Communications Conference, GLOBECOM* [Preprint]. Available at: <https://doi.org/10.1109/GLOBECOM46510.2021.9685899>.

Ghosh, S., Nolte, D., Tsurukawa, F. and Pal R. (2026) "Federated generative imputation with variational and denoising autoencoders for pan-cancer genomics," *Heliyon*, 12(4), p. e44690. Available at: <https://doi.org/10.1016/j.heliyon.2026.e44690>.

Hallaji, E., Razavi-Far, R., Palade, V. and Saif, M. (2021) "Adversarial Learning on Incomplete and Imbalanced Medical Data for Robust Survival Prediction of Liver Transplant Patients," *IEEE Access*, 9, pp. 73641–73650. Available at: <https://doi.org/10.1109/ACCESS.2021.3081040>.

Kazijevs, M. and Samad, M.D. (2023) "Deep imputation of missing values in time series health data: A review with benchmarking," *Journal of Biomedical Informatics*, 144(Mvi). Available at: <https://doi.org/10.1016/j.jbi.2023.104440>.

- Kowsar, I. (2025) "Digital Scholarship @ Tennessee State University DEEP IMPUTATION OF MISSING VALUES USING FEATURE AND." p. 1652. Available at: <https://doi.org/10.11591/ijeecs.v39.i3.pp1652-1660>.
- Kuang, S., Huang, Y. and Song, J. (2025) "Unsupervised data imputation with multiple importance sampling variational autoencoders," *Scientific Reports*, 15(1), pp. 1–16. Available at: <https://doi.org/10.1038/s41598-025-87641-0>.
- Li, X. (2024) "Time Series Forecasting with Missing Data Using Generative Adversarial Networks and Bayesian Inference," *Information (Switzerland)*, 15(4). Available at: <https://doi.org/10.3390/info15040222>.
- Liu, M., Huang, H., Feng, H., Sun, L., Du, B. and Fu Y (2023) "PriSTI: A Conditional Diffusion Framework for Spatiotemporal Imputation," *Proceedings - International Conference on Data Engineering*, 2023-April, pp. 1927–1939. Available at: <https://doi.org/10.1109/ICDE55515.2023.00150>.
- Matrood, H.H. (2026) "Journal of Economics, Business, and Commerce (JEBC)," 3(1), pp. 226–235. Available at: <https://doi.org/10.69739/jebc.v3i1.1980>
- Mattei, P.A. and Freiisen, J. (2019) "Miwae: Deep generative modelling and imputation of incomplete data sets," *36th International Conference on Machine Learning, ICML 2019*, 2019-June, pp. 7762–7772.
- Najeeb, R.S.M. and Dahl, I.O.A.M. (2022) "Brain Tumor Segmentation Utilizing Generative Adversarial, Resnet And Unet Deep Learning," *2022 8th International Conference on Contemporary Information Technology and Mathematics, ICCITM 2022*, pp. 85–89. Available at: <https://doi.org/10.1109/ICCITM56309.2022.10031760>.
- Pereira, R.C., Abreu, P., Rodrigues, P. and Figueiredo M. (2024) "Imputation of data Missing Not at Random: Artificial generation and benchmark analysis," *Expert Systems with Applications*, 249(PB), p. 123654. Available at: <https://doi.org/10.1016/j.eswa.2024.123654>.
- Qasha, R. and Sulyaman, H. (2025) "Microservices caching for container-based IoT system in the edge and cloud," *Indonesian Journal of Electrical Engineering and Computer Science*, 39(3), p. 1652. Available at: <https://doi.org/10.11591/ijeecs.v39.i3.pp1652-1660>.
- Sultan, N.A. and Abdullah, D.B. (2023) "Investigating scientific collaboration networks in Iraq using cloud computing and data mining," *International Journal of Applied Science and Engineering*, 20(4). Available at: [https://doi.org/10.6703/IJASE.202312_20\(4\).003](https://doi.org/10.6703/IJASE.202312_20(4).003).
- Tashiro, Y., Song, J., Song, Y. and Ermon S. (2021) "CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation," *Advances in Neural Information Processing Systems*, 30(NeurIPS), pp. 24804–24816. Available at: <https://doi.org/10.5555/3540261.3542161>.
- Veldkamp, K., Grasman, R. and Molenaar, D. (2025) "Handling missing data in variational autoencoder based item response theory," *British Journal of Mathematical and Statistical Psychology*, 78(1), pp. 378–397. Available at: <https://doi.org/10.1111/bmsp.12363>.
- Wang, D., Yan, Y., Qiu, R., Zhu, Y., Guan, K., Margenot, A. and Tong, H. (2023) "Networked Time Series Imputation via Position-aware Graph Enhanced Variational Autoencoders," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2256–2268. Available at: <https://doi.org/10.1145/3580305.3599444>.
- Wang, X., Zhang, H., Wang, P., Zhang, Y., Wang, B., Zhou, Z. and Wang Y. (2023) "An Observed Value Consistent Diffusion Model for Imputing Missing Values in Multivariate Time Series," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2409–2418. Available at: <https://doi.org/10.1145/3580305.3599257>.
- Xiao, C., Jiang, X., Du, X., Yang, W., Lu, W., Wang, X. and Chetty, K. (2025) "Boundary-enhanced time series data imputation with long-term dependency diffusion models," *Knowledge-Based Systems*, 310, pp. 1–17. Available at: <https://doi.org/10.1016/j.knosys.2024.112917>.
- Xu, X., Hu, L., Liu, J., Miao, X. and Yan X. (2026) "T-LSTM-VAE: A random missing data imputation model for multivariate time series," *Journal of the Franklin Institute*, 363(1), p. 108231. Available at:

<https://doi.org/10.1016/j.jfranklin.2025.108231>.

Yoon, J., Jordon, J. and Van Der Schaar, M. (2018) "Supplementary materials GAIN: Missing data imputation using generative adversarial nets," *35th International Conference on Machine Learning, ICML 2018*, 13(1), pp. 9052–9059.

Zhang, L., Lin, J., Sabatini, J., Zapata-Rivera, D., Forsyth, C., Jiang, Y., Hollander, J., Hu, X. and Graesser A. (2025) "Generative Data Imputation for Sparse Learner Performance Data Using Generative Adversarial Imputation Networks," 18(9), pp. 1–17. Available at: <https://arxiv.org/pdf/2503.18982>.

Zhang, Y. (2024) "Missing Data Imputation for Health Care Big Data using Denoising Autoencoder with Generative Adversarial Network," *Scalable Computing: Practice and Experience*, 25(5), pp. 3850–3857. Available at: <https://doi.org/10.12694/scpe.v25i5.3023>.



Hiba Z. Mohammed holds a Bachelor's degree (2006) in Computer Science. She is currently a Master's student (research phase) in the Department of Computer Science at the College of Computer Science and Mathematics, University of Mosul, in Mosul, Iraq. Her current research interests include data science, data analysis, and artificial intelligence..

E-mail: hiba.25csp78@student.uomosul.edu.iq



Wael Hadeed received his, M.Sc., in 2010, and Ph.D. in 2023 in computer science. He is affiliated with the Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq. His current research interests include data science, data analytics, cloud computing, distributed systems, and operating system.

E-mail: wael.hadeed@uomosul.edu.iq