



A COMPREHENSIVE REVIEW ON MISSING DATA IMPUTATION TECHNIQUES

NoorAldeen Salam Taher*¹ and Nashaat Jasim Mohammed²

^{1,2}Department of Information Management Techniques, Middle Technical University, Baghdad, Iraq.

¹<https://orcid.org/0009-0005-6125-2353>, ²<http://orcid.org/0000-0002-2887-0388>

E-mail: *dac5011@mtu.edu.iq, dr.nashaat@mtu.edu.iq

ARTICLE INFO

Article History

Received: February 05, 2026

Reviewed: March 05, 2026

Accepted: April 01, 2026

Published: June 30, 2026

Keywords:

Missing Data Imputation,
Machine Learning,
Deep Learning,
Generative Adversarial Networks,
High-Dimensional Data,

ABSTRACT

Missing data is a fundamental challenge in scientific research and often leads to biased results and weakened statistical inference. To address this, imputation methods are essential to preserve the original information. This study compares a wide range of imputation techniques categorized into two main types: traditional statistical methods (mean, median, mode, last observation carried forward, regression, multiple imputation), and machine learning and deep learning methods including algorithms such as (k-nearest neighbors, random forests, support vector machines, neural networks, generative adversarial networks). The results indicate that machine learning algorithms, particularly missForest and KNN, consistently provide high accuracy by modeling complex and nonlinear relationships. Furthermore, GAN-based methods such as GAIN, GIMIN, and MisGIMIN represent significant advancements, especially for high-dimensional data with missing rates exceeding 80%. Specifically, the GIMIN algorithm demonstrated superior performance in root mean square error (RMSE) accuracy at high levels of missingness, while MisGAN achieved the lowest FID error. The study emphasizes selecting methods based on the characteristics of the dataset and recommends advanced machine learning algorithms to ensure unbiased inference, while cautioning against simple imputation techniques that may introduce bias.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Missing data refers to information that was supposed to be collected but is not available for certain variables in the data set. This happens when the process of gathering data fails, whether because participants skip questions, sensors malfunction, or systems do not record the values correctly. The absence of this information creates gaps in the data set, which can lead to biased results. We can perform statistical analysis and reduce the accuracy of predictive models for these reasons. Handling missing data properly is essential to ensure reliable and meaningful conclusions [1]. Missing data represents one of the most critical challenges encountered in scientific research. This issue is widespread and directly impacts the availability and completeness of datasets required for generating valid statistical inferences.

Moreover, numerous statistical analysis techniques depend on fully observed data, which places researchers in a position where they must choose between imputing the missing values or excluding records with incomplete information. However, case deletion is generally considered inefficient and often unacceptable, as it results in the loss of original information that may be essential for deriving more reliable and accurate conclusions [2]. Consequently, imputation emerges as the most viable and methodologically sound solution when the goal is to preserve all original observations within the dataset.

Literary studies have discussed many methods for completing missing data, including traditional statistical methods and machine learning methods.

These methods have been used in various fields, such as classification, pattern recognition, prediction, and so on. They have also been applied in the medical, financial, aviation, maritime fields, and more.

II. CLASSIFICATIONS OF METHODS FOR HANDLING MISSING DATA

II.1 TRADITIONAL STATISTICAL METHODS

II.1.1 Imputation With The Mean

It means replacing the missing value with the mean of the observed data, which is a common and simple method used by researchers to handle missing data, but it is considered sensitive to outliers [3].

II.1.2 Completion With The Median

This method involves replacing the missing value with the median calculated from the observed values. Despite the simplicity of this approach, it distorts the data's original distribution and reduces variance, and it is also not suitable for time series data [4].

II.1.3 Mode Imputation

Mode imputation is considered one of the commonly used methods, as it is used with numerical and categorical data. In this method, the missing element is replaced with the most frequent element, and although this method is easy and quick to apply, it may change the statistical distribution of the variable [5]. One of the challenges facing this method is the presence of more than one mode. In this case, the researcher must decide which value to use.

II.1.4 Last Observation Carried Forward

The principle of this method is based on transferring the last observed value to the place of the missing value [4], [6]. This method is widely used in time series data and longitudinal data, where data is measured in the form of repeated time points. Although this method is easy to apply, it produces biased estimates and reduces variance [7].

II.1.5 Multiple Imputation

A good method for estimating missing data, as it generates several different complete datasets by simulating the missing data multiple times to estimate or reflect the degree of uncertainty associated with the missing values. Then, the results of the different datasets are analyzed using statistical methods, and the final results are combined to obtain the best prediction for the missing imputed values [8].

II.1.6 Regression Imputation

Considered one of the directed teaching methods used to estimate missing values in tabular data. This method relies on building a regression model using the observed values [9], where the variable containing a missing value acts as the dependent variable, while the other complete variables are used as independent variables. After training the model, its outputs are used to predict the missing values based on the values of the predictor variables. The most common regression imputation techniques include mean or mode imputation, simple linear regression, multiple imputation, and nonlinear regression.

Although linear and logistic regression are suitable for data with linear relationships and can handle both categorical and continuous data, they may fail to represent complex and nonlinear patterns. The quality of the estimated values depends on the proper selection of independent variables and the efficiency of the model, making the selection process crucial to ensure reliable results. However, this method can sometimes be inefficient because it requires retraining the model each time the locations of missing values within the data change [5].

II.2 MACHINE LEARNING METHODS

II.2.1 K-Nearest Neighbor Imputation

One of the methods of supervised learning [5], which is an effective algorithm for handling missing data in a database. This algorithm relies on identifying a number k of similar samples, where it calculates the distance between the missing sample's information and compares it with the information of similar samples. After determining the closest distance between the two samples, the missing sample's data is replaced with the value of the closest sample [10]. The appropriate distance metric is chosen based on the type of data (Manhattan, Euclidean, Hamming).

II.2.2 Random Forest

It is a popular supervised learning algorithm, which is not limited to building classification and prediction models but is also used in handling missing data. It works by taking multiple sub-samples from the complete data through the bootstrap and random selection of features. Then, a decision tree is built for each sample, and the trees are combined to form the random forest model. Finally, the missing data is input into the model for processing [11].

II.2.3 Support Vector Machines

Are among the commonly used supervised machine learning algorithms and aim to find a decision boundary that achieves the highest separation margin between classes in the data space. The algorithm relies on a set of data points called support vectors, which directly determine the position and orientation of the decision boundary, while completely ignoring points far from the learning process.

The algorithm is characterized by its high accuracy and strong generalization ability, especially in high-dimensional problems or when the data contains limited noise, making it a reliable model in advanced classification and regression applications [12].

II.2.4 Artificial Neural Network

It relies on two main approaches. The first approach resembles supervised learning, where deep neural networks or probabilistic models are used to estimate missing values. The uncertainty resulting from the missing values is integrated into the responses of the neural cells. This approach demonstrates high accuracy close to that of using complete data [13]. The second approach focuses on training networks directly on incomplete datasets. This is done by dividing the data into partially complete subsets, training base networks on these subsets, and then combining them and fine-tuning the final model. This allows handling missing data without the need for complete data during training [14].

II.2.5 Generative Adversarial Networks

Are considered one of the innovative deep learning methods for handling missing data, consisting of a generator that produces new data resembling real data, and a discriminator that distinguishes between real and generated data [15]. In methods like GAIN [16], the generator uses observed data to estimate missing values, while the discriminator evaluates the quality of these estimates using a hint vector that indicates the missingness pattern.

There are developments such as GAMIN [17] and MisGAN [18] that handle very high missing rates and deal with high-dimensional and incomplete data by generating complete data and learning the missingness pattern without prior knowledge of the measurement process. These methods are powerful in estimating missing values, but they may struggle to explicitly handle different data missingness mechanisms, which can sometimes lead to less accurate estimates if additional constraints and optimization mechanisms are not incorporated.

II.2.6 Clustering Based Imputation

It is a method of unsupervised learning aimed at grouping similar elements based on a similarity or distance measure. It is sometimes used to handle missing data by leveraging the structure of the formed clusters. One of its most well-known methods is the K-means algorithm, which starts by selecting random cluster centers and then repeatedly reassigns points to the nearest center until the clusters stabilize, after which the properties of each cluster are used to impute the missing values [19].

However, clustering results depend on several factors, such as the type of distance function, the number of clusters, and the initial positions of the centers, which can affect the quality of the imputation. Some studies [20] have improved the approach by combining clustering with statistical methods or non-parametric kernels, but it still suffers from limitations, the most important of which are the potential propagation of errors and the neglect of certain types of missing data. In general, as data grows and becomes more complex, traditional clustering methods become less capable of handling large and high-dimensional datasets without additional development.

III. EVALUATION METRICS

Methods for measuring algorithm quality are essential tools for evaluating model performance in various fields of machine learning and data analysis. These methods rely on comparing the model's results with actual values in order to determine the model's accuracy and effectiveness. Some of the most important of these metrics are:

- **Accuracy:** It measures the proportion of correct predictions to the total predictions and is often used in classification tasks.
- **Precision:** represents the proportion of correct predictions among all predictions that the model considered positive.
- **Root Mean square Error (RMSE) :** measures the average error after squaring it and then taking the square root, giving more weight to large errors, which makes it sensitive to outliers.
- **Mean Squared Error (MSE):** Calculates the average differences between actual values and estimated values, and is widely used in regression and estimation tasks.
- **Mean Absolute Error (MAE):** measures the average absolute value of the error, and is characterized by being less sensitive to outliers compared to RMSE.

IV. SURVEY OF RELATED WORK

In a study conducted by [3], seven imputation techniques (mean, median, LOCF, KNN, interpolation, MissForest, and MICE) were compared using RMSE and MAE metrics on three datasets (breast cancer, diabetes, heart disease) with missing data rates of 10%, 15%, 20%, and 25%. The MissForest technique consistently showed the lowest errors across most missing data rates in the three datasets. The superior performance of MissForest is attributed to its use of the predictive capabilities inherent in random forests, which excel at modeling complex and nonlinear relationships. MICE came in second in terms of performance, due to its construction of multiple regression models to estimate missing data, generating multiple complete datasets while accounting for uncertainty in the imputation process. The LOCF technique showed the highest errors for the dataset.

Table 1: relative performance of the best methods in two main aspects: interpolation accuracy and classification performance.

Comparison field	Best method	Second best method	Worst method
Imputation accuracy (based on lowest RMSE/MAE)	MissForest	MICE	LOCF (Last Observation Carried Forward)
"Classification performance (Accuracy, Recall, Precision)",	Imputation before feature selection	Feature selection before imputation	-----

Source: Authors, (2026).

The study [4] evaluated methods for imputing missing data in 24-hour particulate matter (PM_{2.5}) concentrations in rural homes, with missing data levels of 20%, 40%, 60%, and 80%. The Markov, random, and mean methods ranked highest in providing the lowest error in estimating 24-hour averages across all levels of missing values. Meanwhile, the Kalman and median filters showed good performance at low to medium levels of missing data (20%-40%).

Table 2: actual numbers (absolute values) of the key performance results (RMSE and MAE) for the best and worst methods across different datasets.

Dataset	Loss Percentage	Method	RMSE	MAE
Breast Cancer	10%	MissForest	≈0.7	≈0.1
Breast Cancer	25%	MissForest (best)	≈1.0	≈0.2
Breast Cancer	average	LOCF (worst)	≈15–20	≈5–7
Diabetes	10%	MissForest (best)	≈6.8	≈1.2
Diabetes	25%	MissForest (best)	≈9.8	≈2.7
Diabetes	average	LOCF (worst)	≈15–20	≈5–7
Heart Disease	10%	MissForest (best)	≈2.4	≈0.5
Heart Disease	25%	MissForest (best)	≈4.5	≈0.9
Heart Disease	average	LOCF (worst)	≈5–6	≈1.3

Source: Authors, (2026).

The study [9] provides a causal analysis of the variation in the performance of classification algorithms when dealing with missing data, focusing on the impact of data characteristics, patterns of missing values, and the imputation methods chosen. The experiment was conducted using six benchmark datasets, three missingness ratios, and three missingness patterns, resulting in an analysis of 226,800 imputed datasets. No method showed consistent overall superiority across all datasets, confirming that determining the best method depends on the interaction with the characteristics of the dataset. In general, the predictive mean matching imputation method showed the best results, while hot-deck imputation was the worst overall. Listwise deletion, on the other hand, saw a significant performance decline with increased amounts of missing data, as the number of deleted records increased. In [5], the study provided a comprehensive summary of technology classifications, loss mechanisms, and evaluation metrics.

Deletion methods (Listwise Deletion, Pairwise Deletion) caused bias and reduced statistical power. Similarly, single imputation methods (Mean, Median, Mode, LOCF) were biased and underestimated variance. In contrast, multiple imputation is considered a reliable option because it restores the natural variance and incorporates uncertainty, providing correct statistical inference. As for machine learning and optimization methods (DL, ANN, GA, PSO, EM), they are less biased and more accurate, especially in nonlinear scenarios and high loss rates. In [10], eight imputation methods (Cart, Simple Regression, EM, MICE, KNN, Cluster, RF) were tested on real cohort study data to predict cardiovascular diseases in China, assuming a 20% missing rate and assessing performance across two dimensions: imputation accuracy itself and subsequent predictive performance. The KNN and RF algorithms achieved the highest predictive performance among the imputation methods, with no significant statistical difference between them in AUC. The study confirms that machine learning-based imputation methods were more effective in capturing relationships within the data, achieving better imputation results and stronger predictive power compared to simple statistical methods.

Table 3: comparison between eight different methods for imputing missing data, in addition to the complete dataset (ALL). MAE and RMSE indicate the accuracy of the imputation (the lower the value, the better the performance), while AUC represents the performance of the predictive model (the higher the value, the better the performance).

Imputation Method	MAE	RMSE	AUC	95% Confidence Interval for AUC	Performance Notes
ALL (Complete Data)	---	---	0.804	(0.796-0.812)	Best performance in prediction
KNN	0.2032	0.7438	0.769*	(0.759-0.779)	Best in imputation accuracy
RF	0.3944	1.4866	0.777*	(0.769-0.785)	Best in prediction among imputation methods
EM	0.6579	2.4939	0.730	(0.719-0.741)	Good performance
CART	0.7183	2.5534	0.726*	(0.715-0.737)	Good performance
MICE	0.8285	2.8699	0.720*	(0.709-0.731)	Good performance
Simple	0.8567	2.9266	0.720*	(0.709-0.731)	Poor performance
Regression	1.0235	3.5548	0.682	(0.667-0.697)	Poor performance
Cluster	1.1383	3.8296	0.668	(0.653-0.683)	Worst performance

Source: Authors, (2026).

As for the study [12], an algorithm was proposed that uses SVM regression to fill in missing values in state attributes. To do this, the roles are reversed, where the missing state attribute is set as the decision attribute, and the original decision attribute is used as one of the state attributes to predict the missing values. The proposed SVM regression method was tested and compared with four imputation methods (minimum distance method, class mean, class median, two closest neighbors mean), and the SVM regression method showed the highest accuracy in filling missing data in the dataset.

Table 4: compares the real values of Mg (Magnesium) and Ca (Calcium) with the estimated values using the proposed method (SVM Regression) and four other regression methods.

Actual value	Method (1): Regression with SVM (proposed method)	Method (2): Mean value for the same category	Method (3): Median value for the same category	Method (4): Average of the two nearest neighbors	Method (5): Nearest neighbor
Mg: 288	287.17	113.4	108	215.5	2112
Ca: 2157	2158.23	2511.1	2220	1882.5	354
Mg: 432	431.80	113.4	2220	202.8	428
Ca: 3870	3869.39	2511.1	108	1546.8	3836
Mg: 166	166.10	113.4	108	209.1	216
Ca: 1806	1483.64	2511.1	2220	1714.6	1624

Source: Authors, (2026).

The study [13] presents a new and theoretically justified mechanism for handling missing data using neural networks, by adjusting the response of neurons in the first hidden layer to match their expected value. Performance was evaluated on three different architectures (Autoencoder, Multilayer Perceptron - MLP, and Radial Basis Function Network - RBFN) and compared with traditional imputation methods and other specialized techniques. The model was tested on recovering the missing parts in MNIST digit images (missing area 13x13), and the proposed method showed the lowest overall recovery error (MSE=0.01056) compared to imputation methods such as KNN (0.01189) and Mean (0.01727), and even compared to Context Encoder (0.01326).

Additionally, the Multilayer Perceptron (MLP) was applied to seizure classification data with random removal of values at high rates up to 90%, where the proposed method clearly outperformed classical imputation methods. The Radial Basis Function Network (RBFN) was also applied to a binary classification dataset containing internally missing values, and the proposed model demonstrated more accurate performance than traditional imputation techniques in all cases, Confirming that using incomplete raw data in neural networks is better than filling in missing features before the learning process.

Table 5: Mean Square Error (MSE) for reconstructing handwritten digits from MNIST, noting that the proposed method ('our') was trained without full data, while the (CE) method used full data during training. A lower value indicates better performance.

Metric	k-nn	mean	dropout	proposed method (our)	CE (Context Encoder)
Total error	0.01189	0.01727	0.01379	0.01056	0.01326
Error inside the mask	0.00722	0.00898	0.00882	0.00810	0.00710
Error outside the mask	0.00468	0.00829	0.00498	0.00246	0.00617

Source: Authors, (2026).

The study addresses [16] new and innovative methods for imputing missing data, known as Generative Adversarial Imputation Networks (GAIN), which are an adaptation of the Generative Adversarial Networks (GANs) framework. Scientific experiments conducted on various datasets showed that the GAIN method significantly outperformed the latest currently used imputation methods. The GAIN model represents a qualitative leap in missing data imputation techniques, as it combines the power of generative adversarial networks with a hint vector guided mechanism to provide high-quality and accurate imputed data.

The study presented [17] advanced methods for imputing missing data based on the Generative Adversarial Networks (GANs) framework: Generative Adversarial Imputation Nets (GAIN) and Generative Adversarial Multiple Imputation Network (GAMIN). GAIN outperformed the latest other imputation methods (MICE, MissForest, Matrix, Auto-encoder, EM) across all tested datasets (Breast, Spam, Letter, Credit, News) in terms of imputation accuracy. As for GAMIN, it is a method specifically designed to handle datasets with a high percentage of missing data (more than 80%) and relies on multiple imputation.

Table 6: measures the root mean square error (RMSE) and the averageb minimum root mean square error for the top k=3 replacements (Top-k RMSE) at different dropout rates.

Missing rate	80%	85%	90%	95%
GAIN	0.250 (0.239)	0.266 (0.253)	0.281 (0.269)	0.307 (0.304)
misGAN	0.255 (0.233)	0.282 (0.256)	0.324 (0.290)	0.368 (0.343)
Ours (GAMIN)	0.224 (0.209)	0.251 (0.234)	0.272 (0.252)	0.306 (0.287)

Source: Authors, (2026).

The study addressed [18] competitive data imputation networks (misGAN) and presents a GAN-based framework specifically designed for learning from incomplete data. The framework was evaluated through a series of experiments on complex, high-dimensional data, demonstrating the effectiveness of the new methodology in handling this type of data.

Table 7: compares the performance of MisGAN against other methods on three different datasets (MNIST, CIFAR-10, CelebA) under two types of data loss patterns (Square Observation and Independent Dropout).

Dataset and Missing Pattern	Loss Rate	Zero Imputation	Mean Imputation	Matrix Factorization	GAIN	MisGAN (Best)
MNIST: Square Observation	0.1 (10%)	≈2	≈2	≈2	≈2	≈0.5
	0.9 (90%)	≈65	≈68	≈28	≈4	≈2.5
MNIST: Independent Dropout	0.1 (10%)	≈2	≈2	≈2	≈2	≈0.5
	0.9 (90%)	≈110	≈105	≈90	≈50	≈5
CIFAR-10: Square Observation	0.1 (10%)	≈60	≈70	≈40	≈20	≈5
	0.9 (90%)	≈150	≈150	≈100	≈80	≈35
CelebA: Independent Dropout	0.1 (10%)	≈280	≈270	≈200	≈260	≈5
	0.9 (90%)	≈310	≈300	≈250	≈200	≈50

Source: Authors, (2026).

The study presents [19] new methods for imputing missing values based on clustering and nearest neighbors, characterized by using newly imputed values to improve subsequent imputation operations. The proposed method was applied to clinical datasets (breast cancer and liver disorder data) and compared with the simple method that does not use imputed values in subsequent imputations. The proposed method showed better performance than the simple method in the clinical datasets tested, but it should be noted that errors occurring in early imputations propagate and amplify through all subsequent imputations.

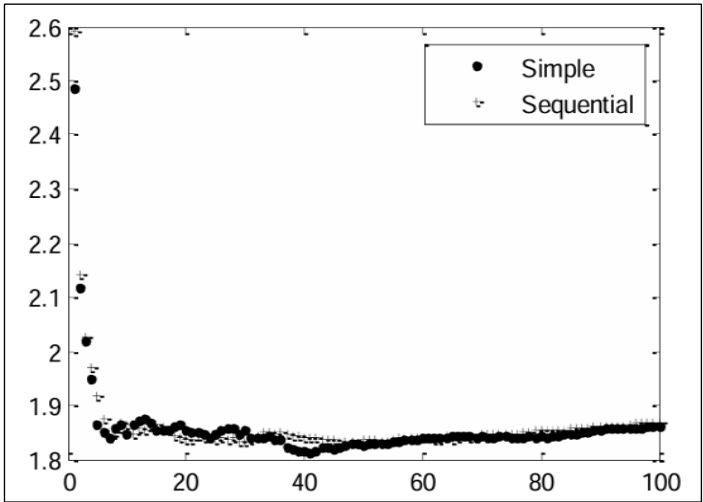


Figure 1: K neighbours versus RMSE for 5% missing data in WBC data.
Source: [19].

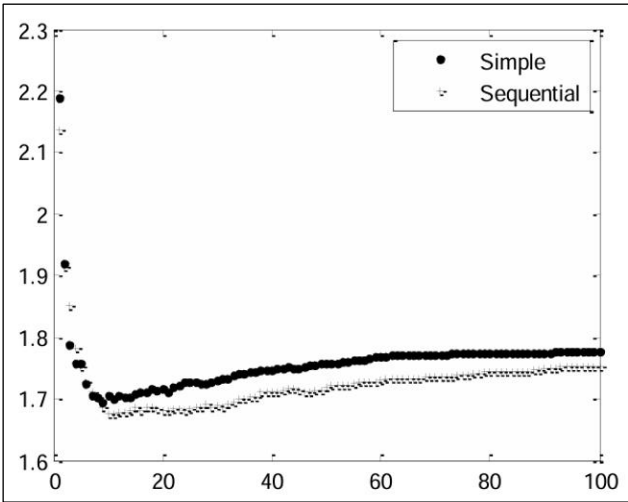


Figure 2: K neighbours versus RMSE for 15% missing data in WBC data.
Source: [19].

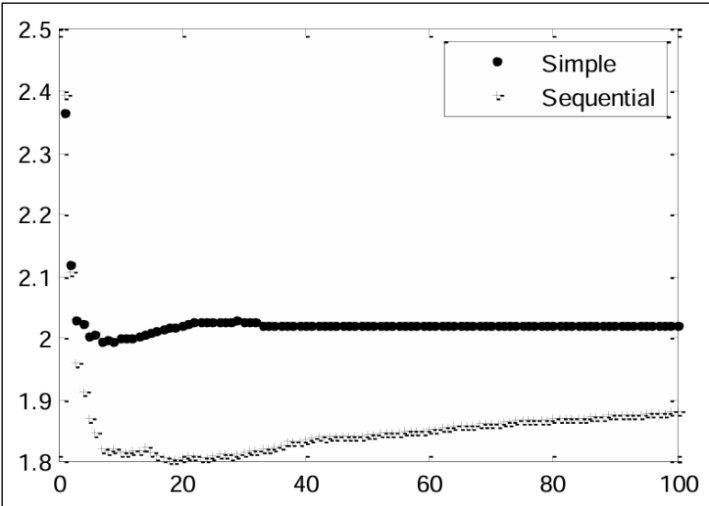


Figure 3: K neighbours versus RMSE for 25% missing data in WBC data.
Source: [19].

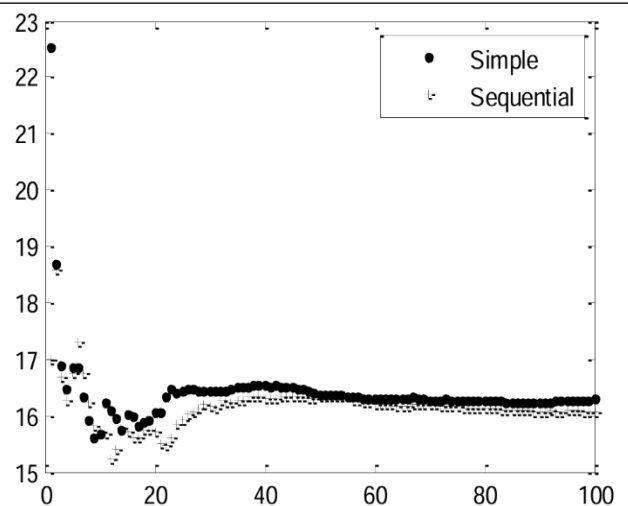


Figure 4: K neighbours versus RMSE for 5% missing data in BLD data.
Source: [19].

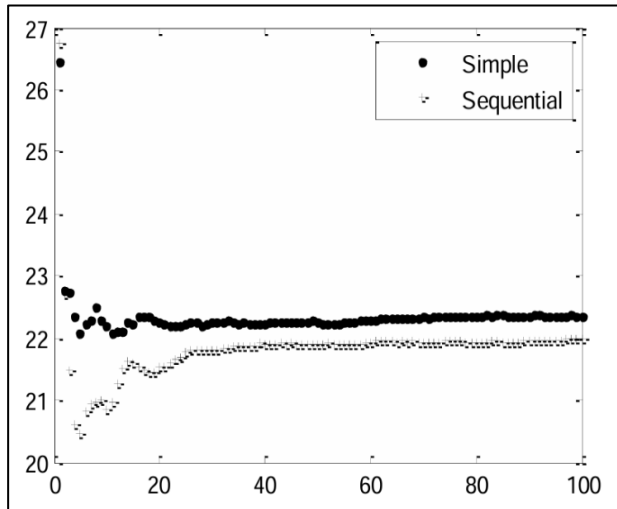


Figure 5: K neighbours versus RMSE for 15% missing data in BLD data.
Source: [19].

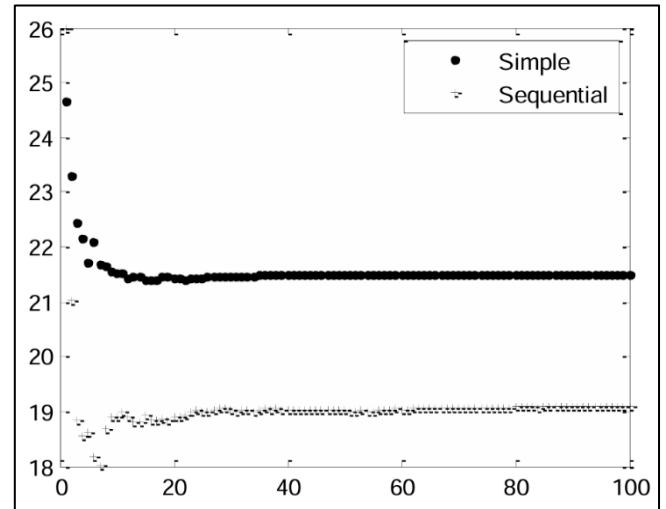


Figure 6: K neighbours versus RMSE for 25% missing data in BLD data.
Source: [19].

V. CONCLUSIONS

Handling missing data is a crucial step to ensure the reliability of statistical inference and the accuracy of predictive models. This comprehensive review has shown that simple statistical methods often lead to biased estimates and reduced data variance, making them a poor choice for complex data. In contrast, multiple imputation provides a reliable alternative that restores natural variance and ensures inference validity, while machine learning algorithms (KNN, MissForest) have demonstrated exceptional capability in capturing complex patterns and nonlinear relationships, offering high retrieval and prediction accuracy. For high-dimensional data with extreme missing rates exceeding 80%, deep learning techniques and generative adversarial networks (GAIN, GAMIN, MisGAN) have made a significant qualitative leap in retrieval accuracy and error reduction compared to other methods. Based on this, and given that no single technique universally outperforms all scenarios depending on data characteristics, the research concludes by recommending the avoidance of simple imputation methods, and instead relying on advanced machine learning algorithms and multiple imputation to ensure the Extracting accurate and unbiased scientific results.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Methodology: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Investigation: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Discussion of results: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Writing – Original Draft: NoorAldeen Salam Taher.

Writing – Review and Editing: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Resources: Nashaat Jasim Mohammed.

Supervision: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

Approval of the final text: NoorAldeen Salam Taher and Nashaat Jasim Mohammed.

VII. REFERENCES

- [1] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [2] C. K. Enders, *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press, 2010.
- [3] L. Oluwaseye, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *Int. J. Data Sci. Anal.*, 2025.
- [4] S. J. Hadeed, M. K. O'Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, "Imputation methods for addressing missing data in short-term monitoring of air pollutants," *Sci. Total Environ.*, vol. 730, Art. no. 139140, 2020.
- [5] L. O. Joel, W. Doorsamy, and B. S. Paul, "A review of missing data handling techniques for machine learning," *Int. J. Innov. Technol. Interdiscip. Sci.*, vol. 5, no. 3, pp. 971–1005, 2022.
- [6] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.
- [7] J. M. Lachin, "Fallacies of last observation carried forward analyses," *Clin. Trials*, vol. 13, no. 2, pp. 161–168, 2016.
- [8] L. J. Beesley, I. Bondarenko, M. R. Elliot, A. W. Kurian, S. J. Katz, and J. M. Taylor, "Multiple imputation with missing data indicators," *Stat. Methods Med. Res.*, vol. 30, no. 12, pp. 2685–2700, 2021.
- [9] J. Sim, J. S. Lee, and O. Kwon, "Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications," *Math. Probl. Eng.*, vol. 2015, pp. 1–14, 2015.

- [10] J. Li et al., "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Med. Res. Methodol.*, vol. 24, 2024.
- [11] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Stat. Anal. Data Mining*, vol. 10, no. 6, pp. 363–377, 2017.
- [12] H. Feng, G. Chen, C. Yin, B. Yang, and Y. Chen, "A SVM regression based approach to filling in missing values," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Springer, 2005, pp. 581–587.
- [13] M. Śmieja, L. Struski, J. Tabor, B. Zieliński, and P. Spurek, "Processing of missing data by neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.
- [14] C.-Y. Cheng, W.-L. Tseng, C.-F. Chang, C.-H. Chang, and S. S.-F. Gau, "A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder," *Front. Psychiatry*, vol. 11, p. 673, 2020.
- [15] I. J. Goodfellow et al., "Generative adversarial networks," in *Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [16] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2018, pp. 5689–5698.
- [17] S. Yoon and S. Sull, "GAMIN: Generative adversarial multiple imputation network for highly missing data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8456–8464.
- [18] S. C.-X. Li, B. Jiang, and B. Marlin, "MisGAN: Learning from incomplete data with generative adversarial networks," *arXiv preprint arXiv:1902.09599*, 2019.
- [19] S. Gajawada and D. Toshniwal, "Missing value imputation method based on clustering and nearest neighbours," *Int. J. Future Comput. Commun.*, vol. 1, no. 2, pp. 206–208, 2012.
- [20] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing value imputation based on data clustering," in *Trans. Comput. Sci. I*, Springer, 2008, pp. 128–138.