IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

THE HIDDEN CHALLENGE: A THOROUGH REVIEW OF MISSING DATA HANDLING AND ITS IMPLICATIONS FOR RESEARCH

[1]Aparna Shukla, [2]Jaya Pal, [3]Rahul Raj

[1]Assistant Professor, [2]Assistant Professor, [3]Student

[1]Department of Computer Science & Engineering,

[1]Birla Institute of Technology, Mesra-Ranchi, INDIA

Abstract: Addressing unavailable data is essential in research to ensure reliable findings. Missing data can introduce bias and undermine validity, arising from participant non-response, measurement errors, or research design flaws. Ignoring or mishandling it leads to poor estimates and inaccurate results. Three main strategies exist removal, imputation, and modelling. Removal discards incomplete cases but can reduce sample size. Imputation replaces missing values with estimates to maintain data integrity. Modelling uses advanced statistical techniques to address uncertainty. The preference of a technique relies on the dataset and the type of missing data.

Index Terms - Missing data, Data analysis challenges, Missing data imputation, Data Integrity.

I. INTRODUCTION

Recent advances in data mining have been significant, but missing data remains a major challenge. Data mining involves analyzing datasets to uncover valuable relationships. However, incomplete data can significantly impact classifier performance and limit information extraction. Regular data collection often results in incomplete datasets, posing issues for statistical analysis and data cleansing. Missing values are common, such as survey respondents omitting income information. Data from multiple sources also frequently suffer from significant information loss[1]. Using incomplete data can lead to inaccurate results, necessitating pre-processing to address anomalies. While small amounts of missing data can be overlooked, substantial gaps must be handled to ensure accurate findings, making pre-processing essential for improving data quality and gaining valuable insights.

Effectively managing missing values is challenging and requires careful examination to identify patterns in the datasets. Since 1980, various solutions have emerged to address these gaps[2]. This document outlines different types of missing values and the methods used to handle them.It's important to differentiate between purged and lost values. Purged values occur when no value can be assigned, whereas lost values indicate an existing but unavailable value in the datasets. Data miners must differentiate between these, as lost data can result from technical issues, conflicts with other data, or entry errors. Understanding the reasons behind missing data is essential before applying any management approach [2], [3].

Missing values can stem from human error, equipment malfunctions, participant non-compliance, withdrawals from research, or merging unrelated data [4][5]. Data-intensive fields often face missing values, leading to reduced performance, processing challenges, and skewed results due to discrepancies between fully populated

and partially populated data [6]. The influence of absent values varies based on factors such as the extent of missing data, the pattern of missingness, and the root causes behind it [7]. A widely-used method to handle missing data is imputation [8], where missing values are substituted with plausible estimates or the instances are removed entirely. Well-known imputation techniques like K-nearest neighbor imputation, mean imputation, and regression imputation are extensively discussed in the fields of statistics and machine learning [9][10].

The remaining part of the paper is structured as follows: Section II explains Missing Data Terminology, Section III describes Strategies of Handling Missing Data. The Evaluation for Imputation of Missing Data Methods is explained in Section IV. Section V gives the comparative analysis of performance metrics and the paper concludes in Section VI.

II. MISSING DATA TERMINOLOGY

Missing data significantly impacts the conclusions of datasets analysis. It occurs at two levels: unit-level, where respondents provide no information (non-response), and item-level, where participants provide partial information [11]. Addressing missing data requires assessing three factors: the amount, the cause, and the distribution of the missing data. Researchers need to assess these elements to determine the most effective approach for managing missing data in their analysis.

2.1 Amount of Missing Data

Data loss undermines statistical conclusions' reliability. No widely accepted criterion for an adequate missing data threshold exists despite extensive experimentation. Surprisingly, this factor has little impact on conclusions compared to others [12][13].
Consider Table 1, which displays patient records from a medical study:

Table 1: Records of Patients

Patient-id	Age	Health status	Contact Details
1	20	Good	Mobile No.
2	Nil	Recovering	Email
3	52	Critical	Nil
4	Nil	Stable	Mobile No.

In this example, the value "Nil" represents missing data. Let's calculate the proportion of missing data for each column:
Age: 2 out of 4 values are missing (IDs 02 and 04), thus, the percentage of missing data for age is 2/4=0.5 or 50**%**.

- **Health Status**: No values are missing.
- **Contact Method**: 1 out of 4 values is missing (ID 03), thus, the percentage of missing data for contact method is ¼=0.25 or 25%.

This example illustrates the percentage of missing data for each column in the dataset.In this instance, missing "Age" data can compromise the study's reliability on aging-related health patterns, while missing "Contact Method" data has less impact on health trend analysis.

2.2 Cause of Missing Data

Missing data, absent values in datasets, arise due to various reasons that include factors like data entry mistakes, equipment failure, low survey participation, deliberate exclusion, technical issues, participant characteristics, survey design flaws, or privacy concerns. Handling missing data poses a common challenge in data analysis, impacting the accuracy and reliability of results significantly.

The matrix data structure can lead to incomplete data in sample analysis. Rubin initially categorized this into three groups based on causes of missing data [14]. Three new types introduced by interestingness algorithms are Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR) [15][16].

2.2.1 Missing Completely at Random(MCAR)

Data is classified, as missing completely at random (MCAR) when its absence is independent of all variables, whether observable or not, implying a random distribution within the matrix X [17]. Yet, this assumption is often impractical in practice. The equation that characterizes Missing Completely At Random (MCAR) is:

$$P(\text{Missing}|X, Y) = P(\text{Missing}) \quad (1)$$

This equation indicates that the likelihood of a value being missing (represented as Missing) is independent of both the observed data X and any unobserved data Y. In other words, the missingness is not related to the values of either the observed or unobserved variables in the dataset.

Survey data includes individuals' favorite colors in Table 2, but some participants left this field blank at random.

Table 2: Favorite Colour Records

Participants	**Favorite Colour**
1	Yellow
2	Purple
3	Nil
4	Green
5	?

In MCAR example, "Nil" symbols denote randomly missing data on favorite colors, showing no pattern or bias when removing rows.

2.2.2. MAR

In Missing at Random.(MAR), the missing data does not exhibit clear patterns. Estimating missing values relies on available information, assuming predictability from other variables, though it may not always fully capture the true relationship [18]. The equation that characterizes Missing at Random (MAR) is:

$$P(\text{Missing}|X, Y, Z) = P(\text{Missing}|X, Z) \quad (2)$$

This equation indicates that the occurrence of missing values depends on observed data and variables Z, but not on unobserved variables Y, establishing conditional independence from unobserved data.To implement MAR, predict missing values using other variables, though this may not capture the full relationship. For example, in a health survey, missing smoking data is more common among lower-income participants.

Table 3: Records of Drug Use

Participant	**Level of Income**	**Smoking Status**
001	High	Smoker
002	Low	Nil
003	Mid	Non-Smoker
004	Low	Nil
005	High	Non-Smoker

In this table, "Nil" indicates missing data, which is more common among low-income participants, showing the MAR pattern.

2.2.3. MNAR

MNAR, also known as "non-ignorable non-response," does not fit into the MCAR or MAR categories. It occurs when the missing value of a variable is influenced by the reason for its absence [19]. The equation characterizing Missing Not At Random (MNAR) is:

$$P(\text{Missing}|X, Y, Z) = P(\text{Missing}|Y) \quad (3)$$

This equation shows that the likelihood of missing data is influenced by the unobserved values Y and cannot be entirely explained by the observed data X. Full information maximum likelihood (FIML) estimation is particularly effective in handling such scenarios [20].

Consider a scenario: Researchers study depression symptoms and medication adherence, collecting data on participants' depression scores and medication adherence. Some participants withhold depression scores, influenced by severity of symptoms (unobserved).

Table 4: Participant's Record

Participant-Id	Depression Score	Medication Adherence
1	20	Yes
2	Nil	No
3	15	Yes
4	Nil	Yes
5	25	No

In Table 4 representing the assumed scenario:

- Participants 2 and 4 have missing depression scores.
- Missingness of depression scores is influenced by severity of depression symptoms (unobserved).

Participants with more severe symptoms are less likely to disclose scores, showing MNAR where missingness relates to the unobserved variable.

2.3 Missing Data Pattern

Missing data patterns reveal the presence and absence of values in a dataset. Although there's no standard classification, [21-23] explore three common patterns: univariate, monotone, and non-monotone..The representation lacks the data schema file, with blue for observed data and red for missing values.

Univariate:

The univariate missing data pattern occurs when only one variable in a dataset has missing values[24]. This pattern is uncommon in many academic fields[25].

Monotone:

The monotone missing data pattern organizes dataset variables sequentially, common in longitudinal studies when participants drop out permanently [26]. Identifying these patterns simplifies data management [27].

Non-Monotonic:

This pattern occurs when missing data in one variable doesn't affect the others [28].

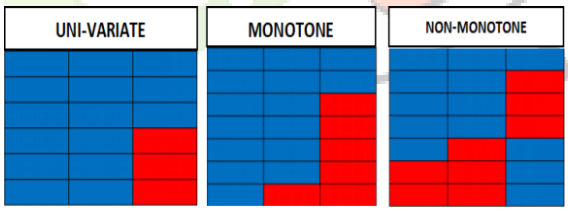

Figure 1: Representation-Missing data patterns

III. Strategies of Handling Missing Data

This section examines many methods for dealing with missing values that have been covered in the literature and how they are used in various contexts [21].

3.1 Ignoring Missing Values (Deletion Methods)

In this method, any entries with missing values are excluded from the analysis, streamlining the process without the need for estimation [18]. However, deletion can introduce bias, especially if data is not randomly missing. Deletion can be pair wise or list-wise [29, 30].

3.1.1. List-wise Deletion

List-wise deletion removes entire cases (rows) with any missing values, for example, in a survey dataset, removing participants with missing responses on any question. If data samples are large and satisfy MCAR, list-wise deletion is reasonable. Smaller samples or violations of MCAR make it less suitable, risking loss of important information with high discarded cases [31, 32].

3.1.2. Pair-wise Deletion

To avoid information loss with list-wise deletion, pair wise deletion retains more data by only omitting values required for specific analyses [33]. However, it can lead to a non-positive definite correlation matrix, limiting coefficient estimation [34]. Pair wise deletion typically produces low bias results for data that is MCAR or MAR [32].

3.2 Imputation Methods

Imputation substitutes missing values with estimated values derived from the available data [35]. Common methods include-

3.2.1. Single Imputation

Simple imputation replaces missing values with the mode, mean, or median of the available values for each characteristic [36]. While widely used for its simplicity, it can introduce bias and unrealistic results in high-dimensional and big data sets [37] [38].

Single-imputation techniques, which usually require less processing resources, seek to compute a correct value for missing data points. Although there is no way to ensure the genuine value, single imputation treats imputed values as real and ignores inherent uncertainty [39].

3.2.2. Regression Imputation

Regression imputation uses full observations to estimate values using a regression model and substitute them for missing data [40]. Regression imputation preserves sample size by retaining observations with missing values but relies on a substantial dataset to ensure robust results. This method utilizes a single regression curve, which may not account for inherent data variability [21]. Using regression, missing values in a feature are predicted from complete attributes, filling in the gaps with estimated values based on available data. Regression technique chosen depends on data nature; for multiple missing features, multivariate regression is used [41]. It assesses linear relationships between multiple predictors and responses [42].Sherwood et. al. [43] used weighted quantile regression to estimate missing health data values, addressing skewness and heteroscedasticity. The approach proved effective for numeric health care cost analysis but relied on fully observed data and was sensitive to missing data rates and functional form specification, potentially introducing bias. In one study, authors used functional principal component regression for missing values, comparing it to complete case analysis. Another study employed multivariate imputation via regression sequences for normal multivariate data, handling multiple missing variables and non-monotonic patterns [44].

3.2.3. Hot-Deck Imputation

Hot-deck imputation matches missing values with complete cases, randomly selecting a donor from similar cases [45]. Another method, nearest neighbor imputation, overlooks missing data variability. Variants include weighted random and sequential hot deck, managing bias by limiting donor selection frequency. This method produces rectangular data, avoids cross-user inconsistency, and doesn't require model fitting, contrasting with parametric methods like regression imputation. It reduces non-response bias but lacks full conceptual clarity despite its research popularity.

Sullivan and Andridge [46] proposed a hot deck method examining MAR to MNAR mechanisms, using fully observed covariates. Simulation studies assessed bias and coverage, showing optimal performance when fully observed values influenced outcomes.

Fractional hot deck imputation was applied to MAR data by Christopher et al. [47], who evaluated its effectiveness against list-wise deletion, mean, and median imputation techniques. Smaller standard errors were produced by their method, which may have outperformed biased imputation techniques in the comparison.

3.2.4. Expectation- Maximization Imputation

The expectation maximization approach uses an iterative procedure of impute, estimate, and repeat until convergence to handle missing variables. Each iteration is split into two phases: expectation, where missing values are estimated from the observed data, and maximization, where these estimates are utilized to maximize the likelihood of the whole data set [48].

Research on handling missing values with expectation maximization (EM) includes Rubin et. al. [49], who studied feeding behaviour in drug-treated and untreated animals. They compared EM to list-wise deletion, Bayesian methods, and mean substitution regression. EM was found to be the most effective, though results may be specific to the dataset's unique characteristics and sampling.

In a separate effort, the problem of training Gaussian mixtures on large, high-dimensional datasets with missing values was addressed by means of an expectation maximization technique for imputation [50]. When compared to simple imputation techniques, the performance of the classification model was greatly enhanced by the imputed datasets. Nevertheless, this method necessitated costly matrix calculations.

Single imputation methods are simple and time-saving for handling missing data. However, they often introduce bias and do not account for the error of their imputations or the uncertainty associated with the missing values. As a result, researchers have developed improved methods that offer better performance and yield unbiased analysis.

3.2.5. Multiple Imputation

Missing data handling goes beyond simple deletion. The drawbacks of single imputation are addressed by multiple imputations, which estimate multiple values expressing uncertainty around the true value [51]. This method creates a single point estimate by averaging the parameter estimates from M samples through a variety of data analysis techniques.

Thus, multiple imputations comprises three phases:
1. Generate M complete datasets by handling missing data.
2. Analyze the M complete datasets.
3. Combine results from the M datasets for the final imputation.

Although multiple imputation is a widely accepted method for managing missing values, researchers need to employ appropriate techniques to ensure dependable results [52]. Because real-world datasets have high percentages of missing values and intricate, nonlinear interactions between variables, imputing them—including survey, clinical, and industrial datasets—can be difficult.
Traditional multiple imputation methods may struggle with high-dimensional data, prompting researchers to refine these algorithms. Moreover, applying techniques designed for continuous data to impute categorical data can lead to biased outcomes [53, 54].

IV. Evaluation Parameters For Imputation of Missing Data Methods

Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Area Under the Curve (AUC) can be used to evaluate how well various approaches handle missing variables.

4.1 Mean Absolute Error (MAE)

The average size of mistakes between expected and actual values is measured by the Mean Absolute Error (MAE), which is computed as the mean of the absolute disparities between them

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (4)$$

where, n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the forecast or imputed value.

4.2. The Mean Squared Error (MSE)

The average squared deviations between the expected and actual values are quantified by the Mean Squared Error (MSE), which is calculated as the mean of these squared differences and assigns a higher weight to larger errors than to smaller ones. Better model performance is indicated by lower MSE values.
MSE defined as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \qquad (5)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted or imputed value, and n is the number of observations.

4.3. The Root Mean Square Error (RMSE)

The Root Mean Squared Error (RMSE) assesses the average size of discrepancies between predicted and actual values. It is computed as the square root of the mean of the squared variances between these values, placing greater emphasis on larger discrepancies.. This is portrayed as:

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

The Mean Squared Error, or MSE, calculates the average squared difference between the actual and expected data. Root Mean Squared Error (RMSE), which yields the standard deviation of these discrepancies, is used to compute the square root of MSE. The estimated missing values are denoted by $\bar{y}_i$), the number of observations is n, and the observed values are represented by yi. A lower score for these performance metrics indicates that the estimated values are comparatively near the true values.

4.4. Area under the Curve (AUC)

The ROC curve, which illustrates the effectiveness of imputation, is summarized and the separability is evaluated by AUC [55]. The true positive rate (TPR) and false positive rate (FPR) serve as its definitions.TPR stands for the percentage of positives that are correctly detected, and FPR for the percentage of negatives that are mistakenly labeled as positives. According to [57], these rates—true positive rate (TPR) and false positive rate (FPR)—are as follows:

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (7)$$

$$FPR = \frac{False\ Positives}{False\ Positives\ +\ True\ Negatives} \quad (8)$$

The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), with their quadratic loss function and ability to quantify forecasting uncertainty, are valuable metrics despite their sensitivity to high values. The MAE, or Mean Absolute Error, on the other hand, is thought to be a more natural and straightforward metric because it is not affected by outliers [58].

Research often uses RMSE for evaluating missing value imputation, despite some arguments favoring MAE for its lower sensitivity to outliers [59]. Chai and Draxler [60] supported RMSE in order to more accurately depict model performance. The AUC, or Area Under the Curve, offers a visual representation of imputation performance and remains unaffected by population distribution and decision criteria but may not reflect actual decision thresholds or model goodness-of-fit. They are not interchangeable, as the discussion over proper measurements demonstrates. The accuracy of each distance measure (MSE, RMSE, MAE, and AUC) is measured in relation to the real non-missing data; the metric used should be in line with the particular analytic needs.

Take a look at the sample in Table 4 for a comparative study of performance measures.Assume that for each given dataset, we have the following actual and expected values:

Actual Values: [4, -1.5, 2, 6]
Predicted Values: [1.5, 0.0, 2, 9]

Table 5: Performance Metrics with Example

Performance Metrics	**Description**
MAE Willmott, C. J., & Matsuura, K. (2005)	1. Determine the absolute disparities between each set of values that are real and those that are expected. $\|4 - 1.5\| = 2.5$ $\|-1.5 - 0.0\| = 1.5$ $\|2 - 2\| = 0$ $\|6 - 9\| = 3$ 1. Sum these absolute differences: 2.5+1.5+0+3=7 2. Divide the sum by the number of of values to get the average: $^{7}/_{4} = 1.75$ MAE=1.75
MSE Chai & Draxler (2014)	1. Calculate the differences : $(4 - 1.5), (-1.5 - 0.0), (2 - 2), (6 - 9) = [2.5, -1.5, 0, -3]$ 2. Square each differences: $[(2.5)^2, (-1.5)^2, (0)^2, (-3)^2] = [6.25, 2.25, 0, 9]$ 3. Calculate the average of these squared differences: $\frac{6.25 + 2.25 + 0 + 9}{4} = 4.375$ MSE= 4.375
RMSE Chai & Draxler (2014)	Take the square root of MSE $\sqrt{\frac{6.25+2.25+0+9}{4}} = \sqrt{4.375} \approx 2.09$ RMSE= 2.09
AUC Bradley (1997)	<u>Let's classify based on a threshold x=2:</u> **<u>Step 1:</u>** Binarize the actual and predicted values based on threshold $Binarize\ Value = \begin{cases} \text{Positive} & \text{if } x \geq 2 \\ \text{Negative} & \text{if } x < 2 \end{cases}$ <u>Binarize Values</u> Actual: [1,0,1,1] Predicted: [0,0,1,1] **<u>Step 2:</u>** Using these binarize values compute TP,FP,TN,FN <u>Actual vs Predicted:</u> 1. TP- Actual: 1, Predicted: 1 2. FP- Actual: 0, Predicted: 1 3. TN- Actual: 0, Predicted: 0 4. FN- Actual: 1, Predicted: 0 **Calculate TPR and FPR:** True Positives (TP): 2 False Positives (FP): 0 True Negative (TN): 1 False Negative (FN): 1 **TPR** = 2/3 **FPR** = 0 **AUC = TPR * (1 - FPR) = 2/3 * 1 = 2/3** Therefore, AUC for this classification scenario with a threshold of 2 is 2/3. AUC =2/3(Threshold=2)

The above table and the associated references provide a concise summary of different performance metrics used for evaluating missing data imputation methods, along with examples for each metric.

V. CONCLUSION

Recent advances in data mining have been significant, but missing data remains a major challenge. Missing values can stem from human error, equipment malfunctions, participant non-compliance, withdrawals from research, or merging unrelated data. Effectively managing missing values is challenging and requires careful examination to identify patterns in the datasets. Thus, ignoring or mishandling it leads to poor estimates and inaccurate results. The three major strategies: removal, imputation, and modeling are explained in this paper to handle missing data. Removal discards incomplete cases but can reduce sample size. Imputation replaces missing values with estimates to maintain data integrity. Modeling employs sophisticated statistical methods to handle uncertainty. The selection of a method hinges on the characteristics of the data and the missing information. Further, a concise summary of performance metrics is explained through example in section IV for evaluating missing data imputation methods.

REFERENCES

1. S. Gupta and M. K. Gupta, “A Survey on Different Techniques for Handling Missing Values in Dataset” International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, vol.4, no.1, pp. 2456-3307, 2018.
2. A. Jadhav, D. Pramod, and K. Ramanathan, "Performance Evaluation of Data Imputation Techniques for Numerical Datasets", Applied Artificial Intelligence, vol. 33, no. 10, pp 913-933, 2019.
3. Suthar B., Patel H., and Goswami A., “A Survey: Classification of Imputation Methods in Data Mining", International Journal of Emerging Technologies and Advanced Engineering, vol. 2, no. 1, pp. 309–312, 2012.
4. R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, “Handling missing data problems with sampling methods,” in 2014 International Conference on Advanced Networking Distributed Systems and Applications. IEEE, 2014, pp. 99–104.
5. O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, “Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry”, Health Quality Life Outcomes, vol. 17, no. 1, pp. 1–9, 2019.
6. Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data. 2nd ed., Wiley Interscience, New York.https://doi.org/10.1002/9781119013563. Ludbrook, J.: Outlying observations and missing values: how should they be handled? Clin. Exp. Pharmacol. Physiol. 35, 670–678 (2008).
7. A. R. Donders, G. J. van der Heijden, T. Stijnen and K. G. Moons, “Review: A Gentle Introduction to Imputation of Missing Values,” Journal of Clinical Epidemiology, Vol. 59, No. 10, 2006, pp. 1087-1091. http://dx.doi.org/10.1016/j.jclinepi.2006.01.014.
8. Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 60, 549576.http://dx.doi.org/10.1146/annurev.psych.58.110405.085530.
9. Baraldi, A. N., & Enders, C. K. , “An introduction to modern missing data analyses”, Journal of School Psychology, vol. 48, no. 1, pp. 5–37, 2010 .
10. Rubin, D.B., “ Multiple Imputation for Nonresponse in Surveys”, John Wiley & Sons Inc., New York, 1987.http://dx.doi.org/10.1002/9780470316696.
11. Little, R. J. A., & Rubin, D. B., “Statistical Analysis with Missing Data”, Hoboken, NJ: John Wiley & Sons, 2002.http://dx.doi.org/10.1002/9781119013563.
12. Zhang, S, Zhang, J, Zhu, X, Qin, Y & Zhang, C.,” Missing Value Imputation Based on Data Clustering”,Transactions on Computational Science (TCOS), vol. 1, pp. 128-138, 2008
13. D.B. Rubin, "Inference and Missing Data", Biometrika, vol. 63, no. 3, pp. 581–592, 1976.
14. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing Value Estimation Methods for DNA Microarrays", Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.
15. L. Xingyi, “Filling missing value algorithm based on Mahalanobis distance and gray analysis”, Journal of Computer Applications, (9), pp. 2502-2506, 2009
16. A. Shukla, S. Kanungo, and V, Bhattacharjee, “Hfna: a Hybrid Folding Neighbour Approach To Handle Missing Data”, International Journal Of Engineering Research & Technology (IJERT), Vol. 12, No. 08, 2023.
17. Chai T, and Draxler RR. “ Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature”. Geosci. Model Dev. Vol. 7, no. 3, pp. 1247–1250, 2014.
18. J. Scheffer, “Dealing with Missing Data", Research Letters in the Information and Mathematical Sciences, vol. 3, pp. 153-160, 2002.

19. S. Singh and J. Prasad, "Estimation of Missing Values in Data Mining and Comparison of Imputation Methods", Mathematical Journal of Interdisciplinary Sciences, vol. 1, no. 2, pp. 75-90, 2013.
20. S. Singh and J. Prasad, "Estimation of Missing Values in Data Mining and Comparison of Imputation Methods", Mathematical Journal of Interdisciplinary Sciences, vol. 1, no. 2, pp. 75-90, 2013.
21. Little R.J., and Rubin D.B., " Statistical analysis with missing data",Hoboken: Wiley, vol. 793, 2019.
22. De Leeuw ED, Hox J., and Huisman, M., " Prevention and Treatment of item Nonresponse", Journal of Official Statistics, vol. 19, no. 2, pp. 153–176, 2003.
23. Berglund P., and Heeringa S.G., "Multiple Imputation of Missing Data using SAS," SAS Institute, 2014.
24. Demirtas H., "Flexible Imputation of Missing Data," Journal of Statistical Software, vol. 85, no. 1, pp 1–5, 2018.
25. Lacerda M., Ardington C., and Leibbrandt M., "Sequential Regression Multiple Imputation for handling incomplete multivariate data using Markov chain Monte Carlo", 2007.
26. Liu, C., "Missing Data Imputation Using the Multivariate t Distribution," Journal of Multivariate Analysis, Elsevier, vol. 53, no. 1, pp 139-158, 1995.
27. Dong, Y., and Peng, CY.J., "Principled missing data methods for researchers".SpringerPlus vol. 2, article 222,2013. https://doi.org/10.1186/2193-1801-2-222.
28. Chen Y-C., "Pattern Graphs: A Graphical Approach to Nonmonotone Missing Data", 2020. https://doi.org/10.48550/arXiv.2004.00744.
29. Patrick E. McKnight, Katherine M. McKnight, Souraya Sidani, and Aurelio José Figueredo, "Missing Data: A Gentle Introduction", New York: Guilford Press; 2007.
30. Graham JW., "Missing Data: Analysis and Design", New York: Springer, pp. 47-69, 2012.http://dx.doi.org/10.1007/978-1-4614-4018-5.
31. Soley-Bori M., " Dealing with missing data: key assumptions and methods for applied analysis.", Boston University, vol. 4, no. 1, pp. 1-9, 2013.
32. Williams R. , "Missing data Part 1: overview, traditional methods", University of Notre Dame; 2015 (last revised).
33. Allison, P. D., " Missing data, Thousand Oaks", Sage Publications, vol. 136, 2001.
34. Kim, J.O. and Curry, J., "The Treatment of Missing Data in Multivariate Analysis", Sociological Methods Research, vol. 6,pp. 215-240, 1977. http://dx.doi.org/10.1177/004912417700600206.
35. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG., Review: A gentle introduction to imputation of missing values", Journal of Clinical Epidemiology, vol. 59, no. 10, pp. 1087–1091, 2006.
36. García Laencina, Pedro J., Sancho-Gómez, José Luis, Figueiras-Vidal, Aníbal R., and Verleysen, Michel, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation", Neurocomputing, vol. 72, pp. 1483- 2009.
37. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martin, M., and Franco, L., "Missing data imputation using statistical and machine learning approaches in a real breast cancer problem", Artificial Intelligence in Medicine., vol. 50, pp. 105-115, 2010.
38. Khan, S.I., Hoque, A.S.M.L. "SICE: an improved missing data imputation technique", Journal of Big Data, vol. 7, article 37, 2020.
39. A. Raghunath, "Survey Sampling Theory and Applications", Cambridge, 2017.
40. Song, Q., and Shepperd, M., " Missing Data Imputation Techniques", IJBIDM, vol. 2, pp. 261-291, 2007.
41. Yu L., Liu L., and Peace K.E., "Regression multiple imputation for missing data analysis", Stat Methods Med Res., vol. 29, no. 9, pp. 2647–2664, 2020.
42. Alexopoulos EC., "Introduction to multivariate regression analysis", Hippokratia, vol. 14, no. 1, pp. 23-28, 2010.
43. Sherwood B., Wang L., and Zhou XH., "Weighted quantile regression for analyzing health care cost data with missing covariates", Stat Med. vol. 32, no. 28, pp. :4967-4979, 2013.
44. Crambes C., and Henchiri Y., "Regression imputation in the functional linear model with missing values in the response", Journal of Statistical Planning and Inference. Vol. 201, pp. 103-119, 2019.
45. Andridge RR., and Little RJ., " A Review of Hot Deck Imputation for Survey Non-response", Int Stat Rev. vol. 78, no. 1, pp. 40-64, 2010
46. Sullivan, D., and Andridge, R.,"A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck", Computational Statistics & Data Analysis, Elsevier, vol. 82(C), pp. 173-185, 2015.
47. Christopher S. Z., Siswantining T. , Sarwinda D., and Bustaman A., "Missing value analysis of numerical data using fractional hot deck imputation.", IEEE 3rd international conference on informatics and computational sciences (ICICoS), p. 1–6, 2019.

48. S.G. Liao, Y. Lin, D.D. Kang, D. Chandra, J. Bon, N. Kaminski, F.C. Sciurba, and G.C. Tsenq, "Missing Value Imputation in High-dimensional Phenomic Data: Imputable or Not, and How?" Bioinformatics, vol. 15, pp. 346-357, 2014.
49. V. Kumutha and S. Palaniammal, "An Enhanced Approach on Handling Missing Values Using Bagging k-NN Imputation", International Conference on Computer Communication and Informatics (ICCCI) in Coimbatore, India, in 2013.
50. D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method Springer Berlin Heidelberg vol. 3066, pp. 573-579, 2004.
51. S. Singh and J. Prasad, "Estimation of Missing Values in Data Mining and Comparison of Imputation Methods", Mathematical Journal of Interdisciplinary Sciences, vol. 1, no. 2, pp. 75-90, 2013.
52. Nguyen CD., Carlin JB., and Lee KJ., " Model checking in multiple imputation: an overview and case study", Emerg Themes Epidemiol, vol. 14, no. 8, 2017.
53. Zhao, Yize, and Qi Long. "Multiple Imputation in the Presence of High-dimensional Data." Statistical Methods in Medical Research, (2013). Accessed June 18, 2024. https://doi.org/10.1177/0962280213511027.
54. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol, vol. 18, no. 1, 2018.
55. Yang, S., & Berdine, G., " The receiver operating characteristic (ROC) curve", The Southwest Respiratory and Critical Care Chronicles, vol. 5, no. 19, pp, 34-36, 2017
56. Gajawada S., and Toshniwal D., " Missing value imputation method based on clustering and nearest neighbours", International Journal of Future Computer and Communication, vol. 1, no.2, pp. 206–208, 2012.
57. .Emmanuel, T., Maupong, T., Mpoeleng, D. et al., " A survey on missing data in machine learning", Journal of Big Data, vol. 8, no. 140, 2021.
58. Y. Kou, C.T. Lu, and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining, pp. 613–617, 2006.
59. Willmott, C., and Matsuura, K., "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance", Climate Research, vol. 30, no. 1, 2005.
60. Chai, T., and Draxler, R., " Root mean square error (RMSE) or mean absolute error (MAE)?", Geoscientific Model Development, 2014.