International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Imputation Strategies for Enhancing Data Preprocessing in Machine Learning

Pushpraj Singh[1], Dr. P. K. Rai[2]
[1]Research Scholar, Department of Computer Science, Awadhesh Pratap Singh University Rewa (MP), India
[2]Head and Professor, Department of Computer Application, Awadhesh Pratap Singh University Rewa (MP), India

Abstract: Data preprocessing is considered a significant step in machine learning and data analysis. During this process, raw data are transformed into a normalized form so that they can be effectively used by machine learning algorithms to produce reliable and accurate results. The quality and completeness of input data show a vital role in effectiveness the performance of Machine Learning (ML) models. However, real-world datasets frequently suffer from missing values due to factors such as human error, equipment malfunction, or incomplete data collection processes. Such missing values can reduce prediction accuracy, introduce bias and decrease model accuracy. This study explores various imputation techniques for effectively handling missing values during data preprocessing. Traditional imputation methods, such as mean, median, and mode, are compared with more advanced techniques, including Multivariate Imputation by Chained Equations (MICE) and K-Nearest Neighbors (KNN). Their performance is evaluated using error metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The experimental results represent that MICE achieved lowest error among the evaluated imputation methods, with MSE, MAE, and RMSE values of 0.082572, 0.421231, 0.649023 respectively.

Keywords: Cervical cancer, Data Preprocessing, Evaluation Metrices, Imputation Techniques, Missing Data, Machine Learning

1. INTRODUCTION

Machine learning (ML) has emerged as an effective tool in healthcare for predictive analysis, helping in the early detection of diseases and supporting personalized treatment decisions. However, the reliability and effectiveness of ML models largely depend on the quality and completeness of the data used for analysis. Medical datasets, including the Cervical Cancer dataset, frequently contain missing values because of patient non-response, recording errors, or incomplete clinical tests. If these missing values are not properly handled, missing data can lead to biased analysis, reduced accuracy, and unreliable predictions.

Cervical cancer is a significant health issue for women worldwide, being the fourth leading cancer globally and continues to contribute significantly to cancer-related deaths. According to the World Health Organization (WHO), approximately 660000 new cases of cervical cancer and around 350,000 related deaths were reported worldwide in 2022 [1]. Early finding of Cervical cancer is essential for significantly reducing death rates. The research focuses on the Cervical Cancer Dataset to explores and compare different imputation techniques for handling missing values and to determine which imputation methods provide the most reliable results. Data cleaning is a fundamental component of data preprocessing and a crucial phase in any data analysis. This involves correcting inaccuracies, resolving inconsistencies, and eliminating errors in the dataset ensuring that the dataset is suitable for analysis and machine learning application.

Unfortunately, real-world datasets are often containing missing values due to various reasons such as human error, incomplete responses, sensor failure, non-response in surveys, or data corruption. These missing values pose a significant challenge, as most ML algorithms are not designed to handle incomplete data and may produce biased or inaccurate results if the issue is not properly addressed.

To overcome this, data preprocessing becomes a life-threatening phase in the machine learning pipeline. One of the most important data preprocessing tasks is **imputation,** which involves estimating the missing values and substituting them with appropriate values in the dataset. Imputation helps to preserve the dataset's structure and statistical properties, enabling the continued use of data that would otherwise be discarded [3].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

The objectives of this research paper are to explore use of various imputation techniques and algorithms to handle missing values in the Cervical Cancer dataset and to compare various imputation techniques and algorithms to determine which one performs best.

1.1 Types of Missing Data-

1.1.1 Missing Completely at Random (MCAR) - **MCAR** refers to a state where missing values occur randomly and are not associated to either the observed or unobserved values in the dataset. To put it another way, the missing values occur purely by chance and are not associated with any other variable or outcome in the dataset. When data are MCAR, the missing values do not introduce bias, and analyses performed on the remaining data yield valid and unbiased results. However, MCAR is a rare condition in real-world datasets, as missingness often depends on observed or unobserved factors. Let X represents the indicator of missingness and Y represents the complete data [2]. The MCAR assumption is mathematically denoted as:

$$P(X \mid Y) = P(X)$$

where, X=1 indicates that an observation is missing and X=0 indicates that it is observed [2].

1.1.2 Missing at Random (MAR) - **MAR** is a missing data mechanism in which the probability of a value presence missing is related to other observed variables in the dataset but not to the missing value itself. In other words, the missingness can be explained by the available data. Unlike MCAR, MAR may introduce bias if it is not handled properly; however, appropriate imputation methods like as multiple imputation, can effectively address MAR scenarios. Let X represents missing values and Y represents observed values. The MAR assumption is mathematically denoted as:

$$P(X \mid Y) = P(X \mid Y)$$

This means that missingness mechanism can be described by using already exited information in dataset [2].

1.1.3 Missing Not at Random (MNAR) - **MNAR** arises when the probability of missingness is directly related to the unobserved value itself, meaning the missing data depends on information that is not available in the dataset Mathematically, MNAR can be denoted as:

$$P(X \mid Y) \neq P(X \mid Y)$$

Thus, probability of missingness mechanism depends on unobserved value or missing value [2].

2. LITERATURE REVIEW

Singaraju, et. al, [3] compares common data preprocessing methods (imputation, scaling, encoding, outlier detection, feature engineering) across various machine learning models and datasets, highlighting the importance of tailored pre-processing strategies for optimal predictive performance.

Meng, H. (2025) discusses various methods for dealing with missing data in machine learning datasets. Common approaches include deletion methods, basic statistical such as mean and median imputations, and advanced imputation methods like as K-Nearest Neighbors (KNN) and multiple imputation. This paper indicates that although deletion methods are easy to implement, they may lead to significant data loss and reduced model reliability. Simple imputation methods are computationally efficient but may introduce bias by altering the statistical distribution of the dataset. On the other hand, machine learning–based methods, like KNN and multiple imputation, can better preserve data relationships and improve predictive performance.

Zhang, S. et al, (2025) evaluated five imputation strategies - GAIN, MICE, Median, MissForest, and MIWA- evaluating their performance across different missing-data mechanisms, using MCAR, MAR, and MNAR, with missingness rates ranging from 10% to 90%. The findings indicate that MissForest and MICE generally performed better than the other methods across most scenarios, indicating their effectiveness in handling missing values under different conditions.

Pindiyan et.al, (2024) evaluated machine learning-based imputation techniques, including KNN, Random Forest, SVM, and Median/Mean Imputation, to address missing data issues by comparing their performance using cross-validation and F1 score matric to identify the most effective strategy for data preprocessing in machine learning applications.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Ayub, H et.al, (2024) addressed the limitations of traditional missing value imputation methods such as KNN, zero, and mean, which fail to capture complex data relationships. It highlights the substantial information loss that occurs when records with missing values are discarded. This research paper proposes an innovative transformer-based approach, which are effective in handling sequential data for imputing missing values.

Ishaq, M. et. al, (2024) explored the use of machine learning techniques for predicting and imputing missing values in categorical datasets. Their study developed ensemble models based on the Error-Correcting Output Codes (ECOC) framework, using SVM and KNN, along with a hybrid model that combines SVM, KNN, and MLP to improve the accuracy of missing-value prediction. The models were evaluated on CPU, Hypothyroid, and Breast Cancer datasets. The results indicates that machine learning-based imputation significantly improves accuracy and robustness, although its performance varied across datasets. ECOC-based ensemble models consistently outperformed individual classifiers in handling missing data.

Aljrees, T. (2024) explored a two-stage framework for the detection of cervical cancer. In the first stage, missing values were handled using KNN-based imputation and normalization techniques. In the second stage, an ensemble-based voting classifier combining Extra Trees, Random Forest, and XGBoost, were applied for prediction. The model achieved high accuracy of 99.41%, demonstrating robust performance in early detection. The study also evaluates PCA-based imputation. A comparative analysis represents that the proposed ensemble approach provides higher accuracy and greater robustness than the existing state-of-the-art.

Sivakani, R. et. al, (2020) described various techniques for dealing with missing values, including mean and median imputation, row-wise and list-wise deletion methods, the KNN, and the Random Forest. Their findings showed that Random Forest approach provided more effective results for handling missing values on large datasets.

3. MATERIALS AND METHOD

This section provides a brief overview of the Cervical Cancer dataset and the various imputation methods used for data preprocessing and handle missing values as well as evaluation metrices.

3.1 Dataset

Cervical-Cancer dataset is used in this study sources from Kaggle. It contains 835 records, 36 features including 35 input variable and an output variable i.e., biopsy outcome which indicates results (cancer positive or negative) [11].

3.2 Missing Value Imputation Techniques

3.2.1 Complete Case Analysis (CCA)

Complete Case Analysis (CCA) is a simple approach for handling missing data. In this method, datasets that do not have complete observations are removed. This approach is also commonly referred to as list-wise deletion. It removes observations with somewhat column that has missing values. This method is one of the simplest and most popular ones in data preprocessing. However, even when the data are Missing Completely at Random (MCAR), using Complete Case Analysis can reduce the size of the dataset and may affect the reliability and performance of machine learning models [4].

3.2.2 Mean Imputation

Mean imputation is an elementary method for handling missing values in numerical data. It replaces each missing value with the average of the observed values in the corresponding feature This method keeps the all observations in the dataset and is efficient in terms of computation, which is why it's widely used in data preprocessing. Although this method is simple and easy to implement, it may introduce bias into the data and can reduce the predictive performance of machine learning models, particularly when dealing with complex datasets [6].

$$Mean = \sum_{i=1}^{n} x_i / n \qquad (1)$$

where x_i denotes observed value, and n denotes numbers of observed (non-missing)

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

3.2.3 Median Imputation

Median imputation is a statistical method that replaces missing numerical values with the median, or middle value, of the observed data in a feature. It works well for datasets that have skewed distributions or outliers because the median is less influenced by extreme values than the mean. This method preserves the central tendency of the data while maintaining the dataset size. While median imputation is straightforward and strong, it does not take into account relationships between variables and may lower data variability. This reduction could impact the performance of machine learning models when the proportion of missing values is high [5].

If n is odd: $mean = \left(\frac{n+1}{2}\right)^{th}$ (2)

If n is even: $mean = [(\frac{n}{2})^{th} + (\frac{n+1}{2})^{th}]/2$ (3)

where, n is the observed value in the attributes

3.2.4 Mode Imputation

Mode imputation is a simple statistical technique used to handle missing values by replacing them with the mode, which is the most frequently occurring value in a particular feature. It is mainly used for categorical or discrete variables, where finding a mean or median doesn't make sense. Mode imputation is straightforward and efficient, and it helps maintain the overall size of the dataset. However, it may increase the frequency of the most common category and reduce data variability, which may lead to bias. Despite these drawbacks, this method is commonly used for preprocessing of categorical data in machine learning and data analysis [6].

$$mode = most\ frequently\ occuring\ value \quad (4)$$

3.2.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbours (KNN) imputation is a machine learning-based technique that determines missing values by identifying observations most similar to the data containing the missing value. The algorithm identifies the K nearest neighbours using the available features and uses their information to estimate the missing value. For numerical variables, the missing value is typically replaced with the mean of the neighboring values, while for categorical variables, the most frequently occurring category among the neighbours is used. KNN imputation can better preserve the relationships between variables and may provide more accurate estimates than simpler methods like mean, mode, and median imputation [10].

$$D_{xy} = \sqrt{weight * squared\ distance\ from\ present\ coordinates} \quad (5)$$

where

$$weight = total\ number\ of\ coordinates/number\ of\ present\ coordinates \quad (6)$$

3.2.6 Multiple Imputation by Chained Equations (MICE)

MICE approach is used for iterative multiple imputation techniques. It is especially appropriate for data with MCAR and MAR missingness mechanism. Each feature of the MICE imputer model with missing values works as a function to other variables. The iterative process of MICE is typically achieved in two stages: initialization and iterative modeling. During the initialization stage, the missing values are replaced with initial estimates, such as mean or median. During the iterative modeling stage, each variable containing missing values is modeled using the other available variables to estimate and update the missing values. This process is repeated many again and again until the imputed values become stable [11].

$$f(x_j | X_{-j}, \theta_j) \quad (7)$$

Where θ_j represents the imputation parameters, and X_{-j} denotes the variables in X excepting x_j.

3.3 Evaluation Metrices

The evaluation metrices are used to asses and compare the performance of different imputation methods. Generally, the imputation method that produce lower MAE, MSE, and RMSE values can be considered the most effective for estimating missing values.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

3.3.1 Mean Absolute Error (MAE) – It evaluates the mean absolute imbalanced between the original values and the imputed values. It provides a straightforward measure of the prediction error without considering whether the error is positive or negative [14].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (8)$$

3.3.2 Mean Squared Error (MSE) – MSE evaluates the mean of the squared differences between the original and imputed values. Since the errors are squared, larger deviations have a greater impact on the final value. Thus, it is very useful when looking for imputation techniques producing large estimation errors [14].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

3.3.3 Root Mean Squared Error (RMSE) - It is obtained by taking the square root of MSE. It provides an overall measure of prediction error while expressing the result in the same units as the original data [14].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

The error performance for five imputation methods is compared and evaluated with MAE, MSE, and RMSE. The methods which produce lowest error indicate the better performance [12].

4. RESULTS AND DISCUSSION

To evaluate the better imputation performance of various imputation methods using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Lowest values of MAE, MSE, and RMSE represents the better imputation performance.

We calculate MAE, MSE, and RMSE values for five imputation methods using python. Libraries such as Pandas, NumPy, and Scikit-learn have used for imputation, and calculation of the evaluation metrics. We have used Cervical Cancer dataset with most appropriate 36 attributes for the study, in which several attributes contain missing values. These missing values are imputed with imputation methods such as mean, median, mode, KNN, and MICE with the help of python programing. These imputation methods are measured with the MAE, MSE, and RMSE and indicates which imputation methods produce the lowest error. The comparative results of various imputation methods are shown in Table 1.

Table 1: Comparison of imputation methods

Method	MAE	MSE	RMSE	Rank
MICE	0.082572	0.421231	**0.649023**	1.0
Mean	0.110022	0.440311	0.663560	2.0
KNN	0.095154	0.475967	0.689904	3.0
Median	0.081667	0.477477	0.690997	4.0
Mode	0.089004	0.5485x19	0.740621	5.0

The experimental results, given in Table-1, represents that Iterative Multivariate Imputation Method that is MICE provide lowest error that compares to others imputation techniques. These experimental results indicate that advanced imputation methods are more effective to maintain the relationship among clinical variable in healthcare dataset.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

MICE achieved better imputation results among all the imputation methods. MICE gain value of MAE, MSE, and RMSE are 0.082572, 0.421231, 0.649023 respectively. The lowest value of MSE, and RMSE represents that MICE produce the very small error among all the imputation techniques.

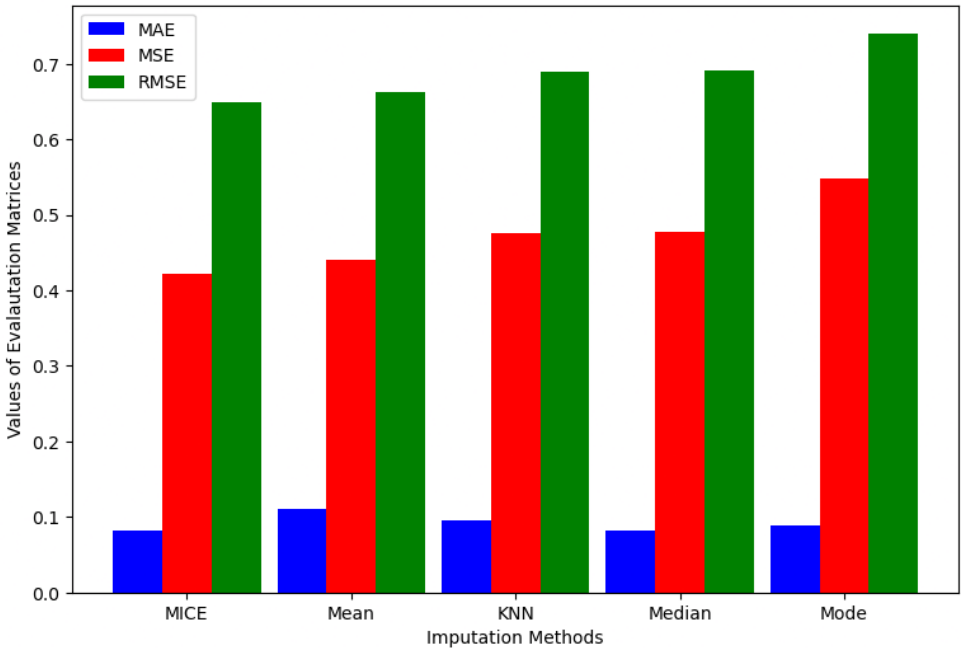
The Mean imputation method obtains the value of MAE, MSE, and RMSE are 0.110022, 0.440311, 0.663560 respectively. In spite of, mean imputation method produces value of MAE relatively high that compares to others imputation methods but the value of MSE and RMSE is closed to MICE imputation method. Thus, it gives better performance to impute the missing values. However, it substitutes the missing values with a central value, so it may reduce the mutability and does not effectively maintain the relationships among cervical cancer attributes.

KNN imputation method achieves the value of MAE, MSE, and RMSE are 0.095154, 0.475967, 0.689904 respectively. Its reconstruction performance is good because the MAE value is relatively low. KNN evaluates missing values using observation with same characteristics. Therefore, it establishes a relationship among attributes. But the value of MSE and RMSE is higher in comparison to MICE and mean imputation method, so it does not give the better performance.

Among all the evaluated imputation methods, median imputation method produces MAE value is 0.081667 which is slightly lower than MICE imputation method. But its value of MSE (0.477477) and RMSE (0.690997) are higher than obtain by MICE. Considering only MAE value, we cannot say that it is best imputation methods. In spite of, median imputation method provides the lowest absolute average error, its higher MSE and RMSE represents that it produces relatively more deviations (errors) for some observations.

Performance of Mode imputation was very poor among all evaluated imputation methods. The value of MAE, MSE, and, RMSE are 0.089004, 0.548519, 0.740621 respectively. higher value of MSE and RMSE indicates that Mode imputation method provides higher deviation between the imputed values and actual values. In general, Mode imputation method is appropriate for categorical variables and is not perfect for numerical variables in any clinical dataset.

Figure 1: Performance comparison among all imputation methods



5. CONCLUSION

Missing values in healthcare dataset can be a serious problem that must be resolved before analysis and classification for disease diagnosis. This study examines both traditional imputation techniques, such as mean, median, and mode, and advanced methods, including KNN and MICE. The study evaluates and compares their performance using different error metrics to identify the most suitable imputation method for handling missing values in the Cervical Cancer dataset. The imputation methods performance is

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

based on three error metrices - MAE, MSE, and RMSE to compare and measure the degree of error in imputed Cervical Cancer dataset. The result shows that MICE imputation method is yielding the lowest degree of error as compare to others imputation methods.

6. REFERENCES

1. https://www.who.int/news-room/fact-sheets/detail/cervical-cancer
2. Sivakani, R., & Ansari, G. A. (2020, September). Imputation using machine learning techniques. In 2020 4th International conference on computer, communication and signal processing (ICCCSP) (pp. 1-6). IEEE.
3. Singaraju, R., Balasubramaniam, K., Agarwal, V., & Jain, A. K. (2025). A Comparative Study of Different Data Pre-processing Methods for Machine Learning. International Journal For Multidisciplinary Research, 7(4). https://doi.org/10.36948/ijfmr.2025.v07i04.52920
4. Meng, H. (2025). A Comparative Study on Missing Value Imputation Techniques in Machine Learning. In SHS Web of Conferences (Vol. 218, p. 02014). EDP Sciences.
5. Zhang, S., Zhang, Z., Zhou, Y., Hong, S., & Liu, H. (2025). Comparison of Imputation Strategies for Incomplete Electronic Health Data. medRxiv, 2025-08.
6. Pindiyan, A. V., & Pramila, R. M. (2024, August). Machine learning-based imputation techniques analysis and study. In 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT) (Vol. 1, pp. 1-6). IEEE.
7. Ayub, H., & Jamil, H. (2024). Enhancing missing values imputation through transformer-based predictive modeling. Igmin Research, 2(1), 025-031.
8. Ishaq, M., Zahir, S., Iftikhar, L., Bulbul, M. F., Rho, S., & Lee, M. Y. (2024). Machine learning based missing data imputation in categorical datasets. IEEE Access, 12, 88332-88344.
9. Aljrees, T. (2024). Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. Plos one, 19(1), e0295632.
10. Mizan, M. T., Ernawan, F., Kasim, S., Erianda, A., & Fauzi, A. M. M. (2025). Comparative Analysis of Robust Imputation Techniques for Enhancing Cervical Cancer Prediction with Missing Data. JOIV: International Journal on Informatics Visualization, 9(5), 1801-1807.
11. Mizan, M. T., Ernawan, F., Kasim, S., Erianda, A., & Fauzi, A. M. M. (2025). Comparative Analysis of Robust Imputation Techniques for Enhancing Cervical Cancer Prediction with Missing Data. JOIV: International Journal on Informatics Visualization, 9(5), 1801-1807.
12. Alade, O. A., R. Sallehuddin, and A. Selamat. "Empirical Performance Evaluation of Imputation Techniques using Medical Dataset." IOP Conference Series: Materials Science and Engineering. Vol. 551. No. 1. IOP Publishing, 2019.
13. https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification.
14. https://33rdsquare.com/evaluation-metrics-with-python-codes/

(This work is licensed under a Creative Commons Attribution 4.0 International License.)