Global Journal of Modeling and Intelligent Computing (GJMIC)

Vol. 3, No. 2, July, 2023 ISSN: 2767-1917

DATA IMPUTATION INSIGHT WITH MACHINE LEARNING IN VARIOUS DOMAINS

Shriya Sahu, Atal Bihari Vajpayee Vishwavidyalaya, India (s.shriya88@gmail.com)
Prerna Verma, Atal Bihari Vajpayee Vishwavidyalaya, India (verma.prerna04@gmail.com)

ABSTRACT

Data imputation has become a critical component in addressing the pervasive issue of missing data across diverse datasets. This paper provides an in-depth exploration of contemporary data imputation techniques, examining their applications, challenges, and emerging trends. It delves into classical imputation methods, such as mean and regression imputation, before surveying advanced approaches like multiple imputation and machine learning-based strategies. The paper investigates challenges inherent to imputation, including handling missing data mechanisms and ethical considerations. Furthermore, the paper concludes the integration of machine learning advancements and the incorporation of domain knowledge into imputation models.

Keywords: Imputation, Missing data, Imputation techniques, Noisy data.

1. INTRODUCTION

Missing or inconsistent data have been a significant challenge in data analysis since the inception of data collection. Addressing this issue has evolved from basic methods like the naive deletion of instances with missing values to more sophisticated modern machine learning imputation techniques. The choice of an appropriate imputation method can be greatly influenced by the mechanism underlying the missing data. There are three primary missing data mechanisms, each of which can impact the accuracy of an imputation method in different ways. The first mechanism is Missing Completely at Random (MCAR), where the missingness occurs entirely at random, with no dependency on any variables within the dataset. In this scenario, the probability of data being missing is the same across all observations, and the missing data is independent of both the observed and the missing values. This implies that the missing data is scattered randomly throughout the dataset without any discernible pattern or reason. The second mechanism is Missing at Random (MAR), where the missingness is related to the observed data but not to the unobserved (missing) data. In other words, the likelihood of data being missing is dependent on the values of the observed data but is independent of the missing values themselves. This mechanism suggests that the missing data can be predicted using the observed data, making it possible to account for the missingness when performing imputations. The third mechanism is Missing Not at Random (MNAR), where the missingness is related to both the observed and the missing values. In this case, the probability of missing data is dependent on the unobserved data itself. This implies that there is a systematic relationship between the missing values and the observed values, which complicates the imputation process because the reason for the missing data is inherently tied to the missing data. Techniques for handling missing values are generally assessed based on their performance under these three missingness mechanisms. Methods must be carefully chosen and validated to ensure they are suitable for the specific type of missing data mechanism present in the dataset. For example, simple deletion might be appropriate under MCAR but could introduce bias under MAR or MNAR. Conversely, advanced machine learning techniques can often accommodate the dependencies in MAR and, to some extent, MNAR, providing more accurate and reliable imputations (Brown, M. L., 2003). The prevalence of missing data is a pervasive challenge across various fields, impacting the reliability and completeness of datasets. In diverse domains such as healthcare, finance, social sciences, and beyond, missing data can arise due to numerous factors, including incomplete survey responses, data entry errors, and sensor malfunctions. The consequences of missing data extend beyond mere numerical gaps; they introduce potential biases, reduce statistical power, and compromise the validity of analyses. In healthcare, for instance, missing patient information can hinder accurate clinical decision-making and impede efforts to develop personalized treatment plans. In financial datasets, missing economic indicators may distort trend analyses and investment strategies. Social sciences heavily reliant on survey data may encounter compromised research outcomes

due to incomplete participant responses. Addressing the challenges posed by missing data is paramount for ensuring the robustness and accuracy of research and decision-making processes. Researchers and practitioners employ sophisticated imputation techniques and statistical methodologies to fill these gaps, allowing for a more comprehensive and reliable understanding of the phenomena under investigation. The prevalence and consequences of missing data underscore the necessity of effective strategies and advanced methodologies to mitigate its impact across a spectrum of disciplines (Brown, M. L., 2003)). The significance of data imputation in maintaining data integrity is rooted in the essential role of complete and accurate datasets for robust analyses and informed decision-making. Without imputation, the integrity of the dataset is compromised, leading to potentially flawed insights and unreliable conclusions. This, in turn, enhances the accuracy and reliability of statistical models, research findings, and subsequent decision-making processes, making data imputation an indispensable step in maintaining the integrity of datasets across various domains.

Addressing missing data is crucial in learning from incomplete datasets, and existing techniques have succeeded with homogeneous attributes, where all independent attributes are either continuous or discrete (Song, Q. et al., 2007) introduces a novel focus on imputing missing data in datasets with heterogeneous attributes, termed imputing mixed-attribute datasets. Despite the prevalence of this scenario in real applications, there is a lack of dedicated estimators for this context. Two consistent estimators for discrete and continuous missing target values propose a mixture-kernel-based iterative estimator tailored for imputing mixed-attribute datasets. Through extensive experiments comparing the proposed method with typical algorithms, the results showcase its superiority in terms of classification accuracy and root mean square error (RMSE) across varying missing ratios, addressing a significant gap in the field of missing data imputation.

The impact of missing data on data mining is profound, influencing the quality and reliability of the extracted knowledge and patterns. Incomplete datasets pose significant challenges to data mining algorithms, as they hinder the algorithms' ability to discern patterns and relationships accurately. The absence of values in certain attributes disrupts the completeness of patterns, potentially leading to biased or skewed results. In data mining tasks such as classification, clustering, and regression, missing data can distort the model's understanding of the underlying structure and relationships within the dataset, compromising the accuracy of predictions and insights. Moreover, missing data can introduce uncertainty, affecting the robustness of decision-making processes based on data mining outcomes. Incomplete information may lead to erroneous conclusions, impacting the efficacy of strategies and interventions derived from data-driven insights. Addressing missing data in data mining requires careful consideration of imputation techniques and methodologies to fill gaps while minimizing bias. The development of robust imputation strategies becomes imperative to mitigate the negative repercussions of missing data, ensuring the integrity of the data mining process and facilitating the extraction of meaningful and reliable knowledge from complex datasets (Zhu, X. et al., 2011).

Bagged tree imputation leverages the principles of bagging (Bootstrap Aggregating) and decision tree algorithms to handle missing data in datasets. The approach involves generating multiple subsets of the original data through bootstrap sampling, where each subset is created by randomly selecting data points with replacement. For each of these subsets, a decision tree model is trained to predict the missing values. The random forest algorithm extends this concept by constructing a multitude of such decision trees, each trained on different bootstrap samples of the data. When imputing missing values, the predictions from all the individual trees are aggregated, typically by averaging in the case of regression tasks or by majority voting for classification tasks. This ensemble method enhances the robustness and accuracy of the imputation process by reducing the variance associated with individual decision trees and mitigating the risk of overfitting. Consequently, bagged tree imputation with random forest effectively captures complex data patterns and dependencies, providing reliable and high-quality estimates for missing values, which is particularly valuable in complex and noisy healthcare datasets (Jordanov et al., 2018).

The classifiers evaluated in the study include Mean Imputation, K-nearest Neighbors (KNN), Fuzzy K-means (FKM), Singular Value Decomposition (SVD), Bayesian Principal Component Analysis (bPCA), and Multiple Imputation by Chained Equations (MICE). SVD decomposes the original data matrix into three constituent matrices: one representing the latent factors of the original features, one diagonal matrix of singular values representing the importance of these factors, and another representing the latent factors of the samples. This decomposition helps in identifying the underlying structure of the data, even when some values are missing. By approximating the missing values based on the observed patterns captured in the decomposed matrices, SVD can effectively fill in gaps in the dataset. The method leverages the correlations among features and samples to provide plausible estimates for the missing entries. This capability is particularly valuable in medical datasets, where missing values are common due to patient non-response or recording errors. While Bayesian Principal Component Analysis (bPCA) is a powerful technique used for imputing missing data in medical datasets by combining the principles of principal component analysis (PCA) with Bayesian inference. Traditional PCA reduces the dimensionality of the data by identifying the

principal components that capture the most variance within the dataset. However, it lacks a probabilistic framework to handle uncertainty and variability in the data effectively. The bPCA addresses this limitation by introducing a Bayesian approach, which estimates the probability distributions of the principal components rather than fixed values. In the context of medical data imputation, bPCA can effectively handle the inherent uncertainties and variations found in medical datasets, such as patient records, clinical trials, and genetic data. By modeling the underlying data distribution, bPCA provides more accurate and robust estimates of missing values. It leverages prior knowledge and observed data to produce a posterior distribution that reflects the range of plausible imputations, thereby improving the quality and reliability of the imputed data. This approach is particularly beneficial in medical research, where precise data imputation is crucial for accurate diagnosis, treatment planning, and epidemiological studies. Each of these methods was applied to four real-world medical datasets: Iris, E. coli, Breast Cancer 1, and Breast Cancer 2. The study meticulously demonstrates the efficacy of these imputation approaches by evaluating their performance across these diverse datasets, which vary in complexity and characteristics. This thorough comparison provides valuable insights into the strengths and limitations of each method, highlighting the contexts in which they are most effective. The findings of this paper contribute significantly to the field of data imputation in healthcare, offering a robust framework for selecting appropriate imputation techniques based on the specific requirements of different medical datasets (Mandel, J. et al., 2015). The Amelia algorithm uses a bootstrapping-based approach to handle missing values. It treats each missing value as a random draw from the observed data distribution and generates multiple imputed datasets. This method captures the variability and uncertainty inherent in the missing data, providing a range of plausible values for each missing entry. By creating multiple complete datasets, Amelia allows for more robust statistical analysis, as the results from each dataset can be combined to produce estimates that reflect the uncertainty due to missing data. Fuzzy Unordered Rule Induction Algorithm (FURIA) is an advanced imputation technique that leverages fuzzy logic to handle missing data in healthcare datasets. FURIA generates fuzzy rules that can manage the imprecision and uncertainty associated with missing values. By using these rules, FURIA can predict missing values based on the patterns and relationships identified within the data. This method is particularly effective for complex medical datasets where relationships between variables are not strictly linear and can be ambiguous. FURIA's ability to deal with such fuzzy relationships makes it a valuable tool for imputing missing data in scenarios like diagnostic data, where symptoms and outcomes might not have clear-cut associations. Multiple Imputation by Chained Equations (MICE) is a flexible and widely used method for imputing missing data in healthcare. MICE works by iteratively imputing missing values for each variable in the dataset using predictive models. Each variable with missing data is regressed on other variables in the dataset, and the missing values are imputed based on these regression models. This process is repeated multiple times to create several complete datasets, which are then combined to account for the uncertainty due to imputation. MICE is highly adaptable and can handle various types of data, including continuous, binary, and categorical variables. Amelia's bootstrapping method provides multiple imputed datasets that reflect the uncertainty in the data. FURIA uses fuzzy logic to handle the complexity and ambiguity of medical data relationships. MICE employs an iterative regression approach to impute values, making it versatile for different types of data and missing patterns. Based on the experiment between three algorithms, MICE performed better to impute real healthcare dataset. (Köse, T. et al.,2020) (Wijesuriya et al., 2020) (Hegde, H. et al.,2021). In IoT systems, missing sensor data can lead to unreliable outputs from ML models, undermining their effectiveness. The data gaps may occur due to various reasons, including sensor malfunctions, connectivity issues, or environmental factors. Consequently, it becomes imperative to handle these gaps effectively to maintain the integrity and reliability of the data used in predictive modeling and decision-making processes. Common practices for dealing with missing data include the complete removal of the affected records or applying simple arithmetic operations, such as mean or median imputation. While these methods are straightforward, they often fail to capture the underlying data distribution and can introduce bias or reduce the variance in the dataset (Okafor et al., 2021). This study investigates the performance of various regression-based machine learning algorithms for imputing missing data in IoT systems. The methods compared include Support Vector Regression (SVR), a robust technique that uses support vector machines to perform regression tasks, effectively capturing non-linear relationships within the data. Decision Tree Regression (DTR) is also examined; it splits the data into subsets based on feature values, making it adept at capturing complex interactions. Ridge Regression is highlighted as an extension of linear regression that incorporates a regularization term to prevent overfitting, which is particularly useful in high-dimensional data. Another method, K-Nearest Neighbors Regression (KNN), imputes missing values based on the values of the nearest neighbors, thereby preserving local data patterns. MissForest (MF) is an iterative imputation method that utilizes random forests, capable of handling mixed data types and capturing intricate data structures. Finally, XGBoost Regression (XGB) is reviewed as an advanced boosting algorithm that constructs a strong predictive model by combining multiple weak models, renowned for its high accuracy and efficiency. Each of these methods brings unique strengths to the table, offering a comprehensive suite

of tools for addressing the challenges of missing data in IoT systems (Aheleroff et al., 2020). The research examines five imputation methods: Variational Autoencoder (VAE), Neural Network with Random Weights (NNRW), Multiple Imputation by Chain Equations (MICE), missForest, and K-Nearest Neighbour (KNN). The analysis reveals that at any given measurement point, auxiliary variables like temperature (T), relative humidity (RH), and other non-missing sensor variables show significant correlations that can be leveraged by imputation methods to predict missing values in target variables (Okafor et al., 2021).

2. CLASSICAL IMPUTATION METHODS

Mean, median, and mode imputation: Mean, median, and mode imputation provide a straightforward way to estimate missing values based on the statistical characteristics of the available data. Mean imputation involves replacing missing values with the mean of the observed values in a specific variable. It is a simple and quick method, suitable for variables with a symmetric distribution. However, mean imputation may be sensitive to outliers, as a few extreme values can disproportionately influence the mean. Median imputation is similar to mean imputation, but it replaces missing values with the median of the observed values. The median is less sensitive to extreme values, making it a robust option when dealing with skewed distributions or datasets containing outliers. Mode imputation involves substituting missing values with the mode, or the most frequently occurring value, in a given variable. This method is particularly useful for categorical variables. However, it may not be applicable if there are multiple modes or if the variable has a continuous distribution. These imputation techniques are simple to implement, but they do not consider the relationships between variables and might not accurately reflect the underlying data structure. Additionally, they can introduce bias and underestimate the variability in the dataset.

Class-based imputation: Class-based imputation is a specialized technique used in the context of missing data imputation, specifically tailored to datasets where the missing values exhibit patterns based on the class or category of the data instances. In this method, missing values are imputed considering the class membership of the instances. This approach recognizes that different classes or groups within a dataset may have distinct patterns of missing data, and imputes missing values accordingly. For example, in a dataset with information on patients grouped by medical conditions (classes), class-based imputation would acknowledge that missing values for specific attributes might follow patterns unique to each medical condition. The imputation process is conducted separately for each class, capturing the inherent characteristics and relationships within that particular group.

Random number imputation: Random number imputation is a data imputation technique used to address missing values in a dataset by replacing them with randomly generated numbers. This method is particularly applicable when the missing data is considered missing completely at random (MCAR), meaning that the probability of an observation being missing is unrelated to its actual value or other variables in the dataset. Random number imputation involves generating random values from the distribution of the observed values in the variable with missing data. The randomness helps preserve the statistical properties of the original data and prevents introducing bias. However, it is important to note that random number imputation assumes the missing data mechanism is MCAR, and its effectiveness may be limited in the presence of systematic missingness or if specific patterns exist in the missing data.

K-NN Distance-Based Median Imputation: K-NN Distance-Based Median Imputation is a technique used to impute missing values in a dataset, leveraging the k-Nearest Neighbors (k-NN) algorithm. In this approach, the missing value for a particular data point is estimated by computing the median of the available values of its k-nearest neighbors in the feature space. The similarity between data points is determined using distance metrics such as Euclidean distance or Manhattan distance. The k-NN algorithm identifies the k-nearest neighbors based on these distances and calculates the median of the observed values for the missing attribute. This method is particularly effective when dealing with continuous or ordinal data, providing a robust imputation strategy that considers the local context of the missing value. However, the performance may be influenced by the choice of the distance metric and the optimal value of k, which need to be carefully determined based on the characteristics of the dataset.

SVM Imputation: SVM imputation is a machine learning-based method that leverages the power of Support Vector Machines. In this approach, a Support Vector Machine is trained on the observed data to learn the underlying patterns and relationships. Once trained, the SVM is used to predict missing values based on the learned patterns, effectively imputing the missing data. SVM imputation is particularly effective when the relationship between variables is complex and non-linear. It excels in scenarios where traditional imputation methods may struggle to capture intricate patterns.

Decision Tree Imputation: Decision Tree imputation involves the use of decision trees, a popular algorithm in machine learning. In this method, decision trees are constructed based on the observed data, mapping relationships between variables. These trees are then employed to predict missing values, utilizing the learned decision rules.

Decision Tree imputation is advantageous for datasets with complex, non-linear relationships, and it provides a transparent and interpretable framework for imputing missing values. The paper introduces two novel techniques for missing data imputation using decision trees and decision forests, namely Decision Tree-based Imputation (DMI) and Splitting and Merging Imputation (SiMI). The methodology of DMI involves using decision trees to identify segments of data where records exhibit higher similarity and attribute correlations. Within these segments, the Expectation-Maximization (EM) algorithm is applied to impute missing values, leveraging the identified similarities to improve accuracy. SiMI extends the approach of DMI by incorporating decision forests to further enhance the segmentation process. It introduces a novel splitting and merging approach that identifies segments with even higher correlations and similarities among records, thus improving the quality of the imputation. The evaluation of these techniques is conducted using nine publicly available datasets, with the performance assessed based on four criteria: coefficient of determination (R^2), index of agreement (d^2), root mean squared error (RMSE), and mean absolute error (MAE). The results from these evaluations indicate that both DMI and SiMI significantly outperform existing imputation methods, such as EMI and IBLLS, particularly in terms of accuracy and statistical reliability. The techniques show substantial improvements in key statistical measures, including confidence intervals (Rahman et al., 2013).

Neural network imputation: Neural network imputation refers to the use of artificial neural networks (ANNs) to estimate and fill in missing values within datasets. Neural networks, inspired by the structure of the human brain, consist of interconnected nodes organized into layers that process information through weighted connections. In imputation, neural networks leverage their ability to capture complex patterns and relationships within data to predict and replace missing values. The network is trained on observed data, learning patterns and correlations, and subsequently used to predict missing values based on the learned relationships. This approach is particularly advantageous for handling non-linear and intricate dependencies in data. Neural network imputation has found applications in various fields, such as healthcare and finance, contributing to more accurate analyses and modeling in scenarios where missing data is prevalent.

The study between various imputation methods for handling missing ordinal data across five datasets. Techniques, including Mean, Class Mean, Median, Class Median, Mode, Class Mode, NN distance-based Median, and classifier-based methods like SVM, decision tree, and neural network, were scrutinized for accuracy and average variance. The findings favored the decision tree imputation, displaying high accuracy and the least variance compared to the original data. Classification experiments using -Nearest Neighbors (NN), Naive Bayes (NB), and Multilayer Perceptron (MLP) algorithms affirmed the positive influence of decision tree imputation on accuracy. Clustering results further highlighted the method's efficacy in closely mirroring clusters derived from the original data, emphasizing its effectiveness in handling missing ordinal data (Alam et al.,2023).

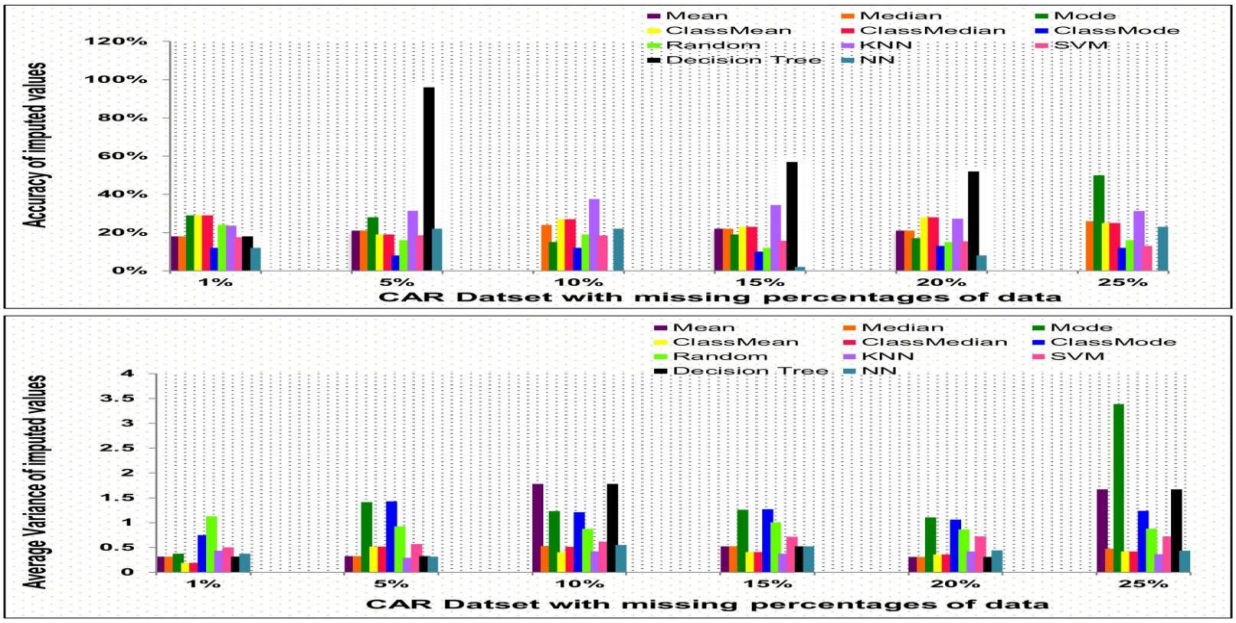

Figure 1 : Accuracy and average variance of various imputation techniques for the CAR dataset (Alam et al.,2023).

Generalized Regression Neural Network (GRNN): The Generalized Regression Neural Network (GRNN) is a type of artificial neural network that excels in function approximation and regression tasks. Developed by Donald Specht, GRNN is characterized by a unique architecture featuring a single hidden layer with radial basis function (RBF) neurons. These neurons perform pattern recognition by calculating the similarity between input patterns and stored prototypes. GRNN is particularly adept at capturing complex non-linear relationships in data, making it suitable for tasks like regression, classification, and function approximation. It boasts a fast learning process during training and provides quick predictions during testing, making it computationally efficient. GRNN's smooth interpolation properties, inherent parallelism, and ability to adapt to varying data distributions contribute to its versatility in diverse applications such as pattern recognition, time-series prediction, and data modeling. The GRNN model's simplicity and effectiveness make it a valuable tool in machine learning and data analysis contexts. Based on GRNN two new models were developed. The proposed nonparametric algorithm, Generalized Regression Neural Network Ensemble for Multiple Imputation (GEMI), addresses incomplete data treatment in data pre-processing. A single imputation version, GESI, was also developed. Evaluated on 98 datasets with varying missing value percentages, GEMI outperformed conventional imputation methods in terms of classification accuracy, interval analysis, and point estimation accuracy. Despite its computational cost and memory storage requirements, GEMI demonstrated superiority over existing algorithms, showcasing its effectiveness in handling missing data and providing a valuable contribution to the field of data imputation (Gheyas et al., 2010).

Another new model using a two-stage soft computing approach for data imputation to assess phishing attack severity. Utilizing a hybrid of K-means algorithm and multilayer perceptron (MLP), missing financial data values are imputed to predict phishing attack severity in financial firms. Post imputation, financial and textual data are mined separately using MLP, probabilistic neural network (PNN), and decision trees (DT). The results demonstrate remarkable classification accuracies of 81.80%, 82.58%, and 82.19% for MLP, PNN, and DT, surpassing prior research. The classifiers also exhibit superior overall classification accuracies across three phishing attack risk levels, highlighting the efficacy of the proposed approach (Nishanth et al. 2012).

A Genetic Algorithm (GA) can be effectively used to determine the optimal parameters for a Fuzzy C-means (FCM) algorithm to impute missing data. The process begins by separately imputing the missing values using Support Vector Regression (SVR) and FCM with initial user-defined parameters. These imputed values are then compared to assess their mutual agreement. If the imputed values are not sufficiently similar, GA is employed to re-estimate the parameters of the FCM. The newly estimated parameters are subsequently used in FCM to impute the missing values again. This iterative process continues until the values imputed by FCM and SVR are similar, ensuring a reliable imputation. GA can enhance the parameter estimation for FCM by artificially introducing missing values into a dataset and then using FCM to impute these values. The difference between the original values and the imputed values serves as a fitness measure in GA, guiding the algorithm to refine the parameters further (Aydilek et al. 2013). The IoT system architecture used in this study comprises multiple interconnected sensors that collect data in real-time. The system ensures data integrity and provides a robust platform for deploying ML models. The architecture includes data preprocessing modules, which handle initial data cleaning and preparation, followed by the application of imputation algorithms to fill in the missing values. To evaluate these imputation methods, missing data in various positions and proportions were introduced into experimentally collected time-series sensor data from a newly developed IoT system platform. This simulated real-world conditions where sensor data might be sporadically missing. The performance of each imputation method was assessed using average Root Mean Squared Error (RMSE) and coefficient of determination (R^2) values. These metrics provide insights into the accuracy and reliability of the imputed data (Kalay, S. et al. 2022).

3. CONCLUSION

In the area of missing data management, various imputation methods are pivotal for maintaining the integrity of datasets. Classical approaches such as mean, median, and regression imputation serve as foundational techniques, providing simplicity and efficiency in handling missing values. Advanced methods, including K-Nearest Neighbors (KNN) imputation and multiple imputation, cater to more complex scenarios, offering nuanced solutions for diverse datasets. Machine learning-based imputation techniques, such as those employing support vector machines (SVM), decision trees, and neural networks, showcase the adaptability of advanced algorithms in imputing missing values with higher accuracy. The paper provides a comprehensive complexity analysis for DMI and SiMI, demonstrating their computational efficiency compared to existing methods. The study further employs various types and amounts of missing data patterns to simulate real-world scenarios and evaluate the performance of the proposed techniques.

These simulations ensure that the techniques are robust and reliable across different datasets and missing data configurations. The results consistently show that DMI and SiMI perform exceptionally well, demonstrating their potential for wide applicability in various domains. The authors suggest that future work should explore the application of these techniques in different fields and with other types of data to further validate their effectiveness and adaptability (Rahman et al. 2013). The study highlighted the critical need for advanced imputation techniques in IoT systems to handle missing sensor data. Among the methods evaluated, Ridge Regression emerged as the most effective, offering robust performance across different missing data scenarios. The experimental results revealed that Ridge Regression outperformed the other ML models for missing data imputation, demonstrating the lowest RMSE and the highest R^2 values. This indicates that Ridge Regression was most effective in capturing the underlying data patterns and providing accurate imputations (Kalay, S. et al. 2022). The study found that the VAE method outperformed other algorithms in imputing missing values, particularly when applied to two real-world datasets: the O3 dataset, which includes three Aeroqual O3 sensors, T, RH, and Federal Equivalent Method (FEM) measurements, and the NO2/O3 dataset, consisting of three Cairclip NO2/O3 sensors, T, RH, and NO2/O3 FEM measurements collected over six months. The dataset with 30% of values imputed using the VAE method was utilized to assess the impact of imputed data on sensor calibration. Calibration models, including Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), and XGBoost (XGB), were trained on the imputed datasets. The results demonstrated that addressing missing values through imputation before sensor calibration significantly enhanced sensor performance, reducing the Root Mean Squared Error (RMSE) between raw sensor outputs and FEM monitor outputs by over 85% (Okafor, N. U. et al., 2021). The choice of imputation method often depends on the nature of the dataset, the extent of missingness, and the goals of the analysis. Classical imputation is favored for its simplicity, while machine learning-based methods excel in capturing intricate patterns and relationships within the data. The impact of imputation extends beyond mere filling of gaps, influencing downstream analyses and decision-making processes. As datasets in various domains continue to grow in complexity, the synergy between classical and advanced imputation methods becomes crucial for researchers, data scientists, and practitioners aiming to extract reliable insights from incomplete datasets. This paper serves as a valuable reference for researchers, data scientists, and practitioners seeking a nuanced understanding of contemporary data imputation strategies and navigating the intricacies associated with missing data in diverse datasets.

REFERENCES

Aheleroff, S., Xu, X., Lu, Y., Aristizabal, M., Velásquez, J. P., Joa, B., & Valencia, Y. (2020). IoT-enabled smart appliances under industry 4.0: A case study. *Advanced Engineering Informatics, 43*, 101043.

Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal, 9*, 100341. https://doi.org/10.1016/j.dajour.2023.100341

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences, 233*, 25–35.

Brown, M. L. (2003). Data mining and the impact of missing data. *Industrial Management and Data Systems, 103*(8), 611–621.

Gheyas, I. A., & Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing, 73*(16–18), 3039–3065. https://doi.org/10.1016/j.neucom.2010.06.021

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked, 17*, 100275.

Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers accuracy improvement based on missing data imputation. *Journal of Artificial Intelligence and Soft Computing Research, 8*(1), 31–48.

Kalay, S., Çinar, E., & Sarıçiçek, İ. (2022). A comparison of data imputation methods utilizing machine learning for a new IoT system platform. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)* (Vol. 1, pp. 69–74). IEEE.

Köse, T., Özgür, S., Coşgun, E., Keskinoğlu, A., Keskinoğlu, P., & Mrozek, D. (2020). Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. *Biomedical Research International, 2020*.

Mandel, J., & Mandel, S. P. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics, 6*(1), 1–6.

Nishanth, K. J., Ravi, V., Ankaiah, N., & Bose, I. (2012). Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. *Expert Systems with Applications, 39*(12), 10583–10589.

Okafor, N. U., & Delaney, D. T. (2021). Missing data imputation on IoT sensor networks: Implications for on-site sensor calibration. *IEEE Sensors Journal, 21*(20), 22833–22845. https://doi.org/10.1109/JSEN.2021.3105442

Rahman, M. G., & Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowledge-Based Systems, 53*, 51–65.

Song, Q., & Shepperd, M. (2007). A new imputation method for small software project data sets. *Journal of Systems and Software, 80*(1), 51–62. https://doi.org/10.1016/j.jss.2006.05.003

Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., & Lee, K. J. (2020). Evaluation of approaches for multiple imputation of three-level data. *BMC Medical Research Methodology, 20*(1), 1–15.

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2011). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering, 23*(1), 110–121. https://doi.org/10.1109/TKDE.2010.99