JOIV : Int. J. Inform. Visualization, 8(3) - September 2024 1271-1276

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Comparative Analysis of Imputation Methods for Enhancing Predictive Accuracy in Data Models

Nurul Aqilah Zamri [a], M. Izham Jaya [a,*], Indrarini Dyah Irawati [b], Taha H. Rassem [c], Rasyidah [d], Shahreen Kasim [e]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia*
[b] *School of Applied Science, Telkom University, Bandung, Indonesia*
[c] *School of Computer Science and Informatics, De Montfort University, Leicester, United Kingdom*
[d] *Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia*
[e] *Soft Computing and Data Mining Centre (SMC), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia*

*Corresponding author: *izhamjaya@umpsa.edu.my*

***Abstract*—The presence of missing values within datasets can introduce a detrimental bias, significantly impeding the predictive algorithm's ability to discern patterns and accurately execute prediction. This paper aims to elucidate the intricacies of data imputation methods, providing a more profound understanding of prevalent imputation methods, including list-wise deletion (IGN), mean imputation (AVG), K-Nearest Neighbors (KNN), MissForest (MF), and Predictive Mean Matching (PMM). The dataset employed in this study consists of financial data about S&P 500 companies in the Compustat North America database. The training and validation dataset encompasses 1973 instances, consisting of data during the fourth quarter of 2009, the first quarter of 2010, and the third quarter of 2014. Within this set, 457 missing values were identified and imputed. The test dataset comprises 197 randomly selected instances from the fourth quarter of 2014, equivalent to ten percent of the total instances in the training dataset. The evaluation findings prominently position the dataset derived from MF imputation as the leading performer among all the imputed datasets. The insights derived from this study are intended to assist practitioners in making informed choices when selecting the most suitable data imputation method, particularly in the context of predictive modeling tasks.**

***Keywords*—Missing value; imputation; predictive modeling; machine learning.**

Manuscript received 15 Jan. 2024; revised 9 Jun. 2024; accepted 19 Aug. 2024. Date of publication 30 Sep. 2024.

I. INTRODUCTION

Missing values refer to undesirable null entries within a dataset and are closely intertwined with the measurement of the data completeness dimension. The occurrence of missing values can be attributed to a broad spectrum of data-related activities encompassing the data collection, including document reviews, interviews, and questionnaires. Furthermore, it is worth noting that missing values can also be unintentionally introduced during experiments, as recorded data may be subject to omissions resulting from equipment malfunction, human errors, and faulty data transmission. These factors are inherently stochastic and pose significant challenges in terms of control. The issue of missing values is inescapable, persisting even when conscientious measures have been implemented during any data-related endeavor.

Within a decision-making context, predictive models empower decision-makers to anticipate the outcomes of their choices. To make accurate predictions, these models require a complete dataset, which forms the basis for constructing a statistical model. Scholars [1]–[3] have emphasized the adverse consequences of utilizing datasets with unaddressed missing values, including diminished predictive model accuracy and biased prediction outcomes.

The abovementioned issue extends to predictive algorithms such as neural networks, where missing values in the training dataset can introduce bias, adversely affecting pattern learning and prediction performance [4]. Remarkably, the situation worsens with many missing values, potentially leading to inaccurate decision-making. Given the paramount importance of accuracy in predictive modeling, it becomes imperative to employ appropriate missing values imputation

methods to estimate and replace missing values. However, before implementing any imputation method, a thorough analysis is essential to comprehend the missing values' patterns, extent, and underlying mechanisms, thereby facilitating the rectification of their root causes.

The objective of this paper is to enhance comprehension of data imputation methodologies and conduct a comprehensive assessment of prevalent imputation methods, including list-wise deletion (IGN), mean imputation (AVG), K-Nearest Neighbors (KNN), MissForest (MF), and Predictive Mean Matching (PMM). The insights derived from this study are intended to assist practitioners in making informed choices when selecting the most suitable data imputation method, particularly in machine learning tasks involving financial datasets.

Missing values can be classified into three distinct mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). Within the MCAR category, the occurrence of missing values is entirely independent of any other observed values or the data of interest itself. Importantly, imputation methods that address MCAR-type missing values do not introduce bias into subsequent analyses [5]. Nevertheless, exercising caution is necessary when testing MCAR patterns, as such patterns are infrequently encountered in datasets. Little's MCAR test can be conducted to confirm the MCAR assumption empirically.

Conversely, missing values fall into the MAR category when their occurrence depends on other observed values but not on the missing value itself. [6] and [7] have elaborated on the feasibility of estimating missing values with a MAR pattern by utilizing other observed values, given their dependency on non-missing values. In contrast, NMAR emerges when the likelihood of missing values in the dataset is associated with unobserved values. Consequently, missing values within the NMAR framework cannot be estimated using other observed values [8].

Imputation methods such as IGN delete cases with missing values, and the subsequent data analysis process is conducted only on cases with complete data. IGN is straightforward, but it reduces the size of instances. As the size decreases, the data analysis process will retain statistical power and avoid difficulties in discovering minor effects or relationships between variables involved in the analysis [4]. Additionally, [4] also explained that adopting IGN in a dataset with an MCAR mechanism increased the standard errors and decreased the significance level in the analysis.

Batista and Monard [9] advocated the application of the KNN algorithm for missing value imputation, relying on the similarity distance between missing values and their K-nearest neighbors, where the data values are available. Typically, well-established distance functions like Euclidean, Manhattan, and Pearson are employed to quantify the proximity between the neighbors and the missing value. Enhancement of KNN iterations has been proposed by using other distance functions, such as the Gray Relational Grade (GRG) [10], and its enhanced version, Reduced Relational Grade (RRG) [11], can be utilized. However, when considering distance functions for assessing attribute similarity, a study by [12] has cautioned that KNN imputation is most effective for datasets exhibiting a robust correlation between attributes. Alternatively, another study by [13] integrates compressive sensing and KNN methods to complete the missing internet traffic matrix. This underscores the importance of assessing the suitability of KNN imputation in the context of the dataset's attribute interrelationships.

MF, a machine learning-based imputation method, is designed to enhance imputation accuracy. MF is an iterative adaptation of the random forest algorithm initially proposed by [14]. This method leverages a random forest model to predict missing values. The procedure begins by estimating the missing values, often through mean imputation or other established imputation techniques. Subsequently, a random forest model is generated for each missing value to facilitate imputation. This process iterates until specific stopping criteria are satisfied, typically triggered by a substantial disparity between the prior and the newly imputed data matrices.

The principal advantage of MF lies in its capability to handle diverse data types while consistently delivering superior imputation results compared to alternative methods. A comparative analysis was conducted on MF and KNN imputation across continuous and categorical variables, demonstrating that MF consistently outperforms KNN, with performance disparities widening as the extent of missing data increases within the dataset [15]. Similarly, in a comparative study by [16], MF was evaluated against other machine learning-based imputation methods, and the findings underscored MF's superior imputation accuracy and computational efficiency. However, it is noteworthy that, akin to other machine learning-based approaches, the iterative nature of the imputation process renders MF computationally intensive, mainly when applied to large datasets.

In the context of AVG, the approach entails replacing missing values with the mean of the marginal distribution derived solely from the available data points. Importantly, this imputation method does not consider any conditional relationships with other variables that may be associated with the missing value. Consequently, the introduction of mean imputation disrupts the inherent data randomness, as missing values are substituted with a fixed constant, thus negating the potential variability in the data values. Veering from data randomness during imputation can lead to compromised statistical inference [17]. The alteration of randomness within imputation procedures can only be circumvented by imputing non-constant values in datasets containing missing values. This ensures that the variability and complexity of the original data are preserved, aligning with robust statistical practices.

PMM, as introduced by Little in 1988, employs linear regression to model variables with missing values based on the set of variables without missing values within the dataset. This method entails deriving a set of coefficients through regression analysis, which is subsequently utilized to predict values for the dataset. The predicted value closest to the observed value is then chosen as the donor value to replace the missing data point. PMM inherits a "hot deck" characteristic by leveraging non-missing data values from the same dataset to impute the missing values. Consequently, PMM ensures that imputed values remain within the range of observed values in the dataset. For instance, if the observed variable contains only positive numbers, PMM guarantees that the imputed values remain consistently positive. In a

study by [18] evaluating PMM's performance for semi-continuous data in comparison to multiple imputation, the results demonstrated that both imputation methods performed equivalently, with no significant differences in imputation bias observed. Furthermore, the study established that PMM preserves the original data distribution during the imputation process, ensuring that the imputed values never extend beyond the range of observed values. This underscores PMM's capacity to maintain the integrity of the dataset's underlying distribution.

II. Material and Method

This study employs a systematic approach to ensure a comprehensive and rigorous analysis of the prediction model. This section elucidates the approach undertaken, as illustrated in Fig. 1.

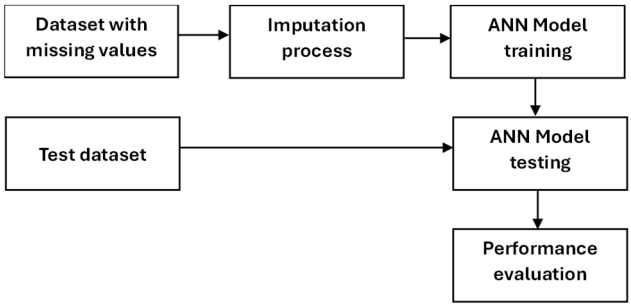

Fig. 1 Systematic approach in the study

The dataset utilized in this study consists of financial data pertaining to S&P 500 companies, which was sourced from the Standard & Poor's Compustat North America database. Compustat is a comprehensive repository of financial, statistical, and market information encompassing active and inactive companies across the global landscape. The choice of the Compustat database in this research is predicated on its extensive utilization in data quality research, particularly studies involving financial data. Specifically, this study extracted financial data for S&P 500 companies during the fourth quarter of 2009, the first quarter of 2010, the third quarter of 2014, and the fourth quarter of 2014 from the Compustat database. The dataset extracted fifteen financial variables, enabling the construction of fourteen financial ratios, as initially proposed by [19]. The resulting financial ratios are presented in Table 1.

TABLE I
Financial ratio and financial variables in the dataset collection

Financial Ratio	Financial Variables
Earnings before interest and taxes/total assets	(Net Income (Loss) (+) Interest and Related Expense – Total (+) Income Taxes Payable) ÷ (Assets – Total)
Net income/net worth	Net Income (Loss) ÷ (Stockholders Equity – Total)
Gross profit/total assets	(Sales/Turnover (Net) (–) Cost of goods sold) ÷ (Assets – Total)
Net income/gross profit	Net Income (Loss) ÷ (Sales/Turnover (Net) (–) Cost of goods sold)
Current liabilities/total assets	(Current Liabilities – Total) ÷ (Assets – Total)
Total liabilities/total assets	(Liabilities –Total) ÷ (Assets – Total)
Long term debt/total equity	(Long-Term Debt – Total) ÷ ((Assets-Total) – (Liabilities-Total))
Current assets/current liabilities	(Current Assets – Total) ÷ (Current Liabilities – Total)
Inventories/current liabilities	(Inventories – Total) ÷ (Current Liabilities – Total)
Interest expenses/sales	(Interest and Related Expense- Total) ÷ (Sales/Turnover (Net))
Selling, general & administrative expenses/sales	(Selling, General and Administrative Expenses) ÷ (Sales/Turnover (Net))
Accounts receivable/sales	(Receivable – Total) ÷ (Sales/Turnover (Net))
Accounts payable/inventories	(Account Payable/Creditors – Trade) ÷ (Inventories – Total)
Accounts payable/sales	(Account Payable/Creditors – Trade) ÷ (Sales/Turnover (Net))

Among these ratios, two financial variables, Earnings per Share and Dividends per Share, are employed to compute the RCSE class labels. Subsequently, the RCSE values transform three-valued class labels by applying the equal-width method. Specifically, these labels are categorized as follows: DOWN (RCSE $\leqslant$ 0.011), NOCHG (RCSE $\leqslant$ 0.104), and UP (RCSE > 0.105), by the approach outlined by [19].

The dataset collection is divided into two sets. The first set encompasses 1973 instances, consisting of data from 500 S&P companies during the fourth quarter of 2009, the first quarter of 2010, and the third quarter of 2014. Within this set, 457 missing values were identified. During the evaluation phase, these missing values underwent imputation through various methods, including AVG, IGN, PMM, KNN, and MF, yielding ten distinct datasets as outcomes. Subsequently, the prediction model utilized these imputed datasets for training and validation purposes.

The second set comprises 197 instances, equivalent to ten percent of the total cases in the first dataset. These instances were randomly selected from the fourth quarter of the 2014 dataset, with missing values excluded. This second dataset serves as the testing set for evaluating the prediction model's performance, as detailed in Table 2 provides an overview of the datasets used in this research.

TABLE II
Dataset collection

Set	Number of datasets	Usage	Number of instances	Missing values
1	10	Training Validation	1973	457
2	1	Testing	197	0

Notably, the dataset size adheres to the common practice in designing neural networks for financial data, fixed at 70% of

the total instances in the training and validation dataset. Additionally, an early stopping strategy is implemented, wherein training ceases in the absence of improvements in generalization error or upon reaching the maximum training time. To assess the classifiers' performance and test the models' generality, a 10-fold cross-validation approach is employed. Moreover, the size of the validation dataset is set at 20% of the total instances in the first set. The out-of-sample testing dataset encompasses 10% of the total cases in the training and validation dataset. Instances within this dataset are randomly selected from the fourth quarter of the 2014 Compustat dataset, excluding those with missing values. During the testing phase, the trained Artificial Neural Network (ANN) classifiers are employed to predict RCSE values for each company within the testing dataset. Subsequently, the trained classifiers are evaluated to validate the prediction model's performance.

The evaluation phase comprises five experiments in which datasets with missing values undergo imputation utilizing OFFDM, AVG, IGN, PMM, KNN, and MF methods. These imputed datasets are subsequently employed to train the prediction model for forecasting RCSE class values within the testing dataset. The prediction model's performance is assessed using three key metrics: root mean squared error (RMSE), prediction accuracy, and F-measure. These metrics are quantitative indicators to gauge the prediction model's effectiveness when various imputation methods address missing values within the training dataset [20].

RMSE quantifies the dissimilarity between predicted values generated by the model and the actual observed values within the dataset, which plays a pivotal role in this context. A diminished RMSE value signifies a reduced error rate and a more precise and accurate prediction. The calculation of RMSE is formally presented in Equation (1), encapsulating its quantitative essence in assessing the predictive performance of the models where e_{ori} is the observed value, e_{est} is the predicted value by the model and is the total number of predictions.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_{ori} - e_{est})^2} \quad (1)$$

Prediction accuracy is a critical metric that quantifies the proportion of target class values correctly forecasted by the prediction model. This metric, derived from the information within the confusion matrix, involves computing the sum of true positive (TP) and true negative (TN) values, which is then divided by the total number of instances within the test dataset. TP denotes instances predicted as positive that are indeed positive, while TN represents instances predicted as negative that are genuinely negative. The calculation for prediction accuracy is formally presented in Equation (2), serving as an essential gauge of the model's ability to make correct predictions.

$$Prediction\ Accuracy\ (\%) = \frac{tp+tn}{tp+tn+fp+fn} \times 100 \quad (2)$$

The F-Measure, alternatively recognized as the harmonic mean of recall and precision, is computed utilizing Equation (3). This metric encapsulates the balance between recall and precision, providing a comprehensive assessment of model performance.

$$F - Measure = \frac{2\ X\ Precision\ X\ Recall}{Precision+Recall} \quad (3)$$

Precision measures the percentage of accurate results among all the predicted results, while recall quantifies the percentage of correctly predicted results within the entirety of predicted results. The F-Measure attains higher values when both recall and precision exhibit excellence. Recall is determined by the formula (TP / (TP + FN)), where TP represents true positives, and FN stands for false negatives. Precision is computed as (TP / (TP + FP)), with TP denoting true positives and FP representing false positives.

III. Results and Discussion

Table 3 provides the RMSE outcomes for datasets subjected to imputation through PMM, MF, KNN, AVG, and IGN techniques.

TABLE III
RMSE BETWEEN IMPUTATION METHODS

Dataset	RMSE
MF	0.4536
PMM	0.4579
AVG	0.4584
KNN	0.4605
IGN	0.4845

The results notably highlight the dataset's inferior RMSE performance under IGN imputation compared to alternative imputation methods. Fig. 2 visually illustrates the discernible RMSE discrepancies between IGN and the other imputation methods.

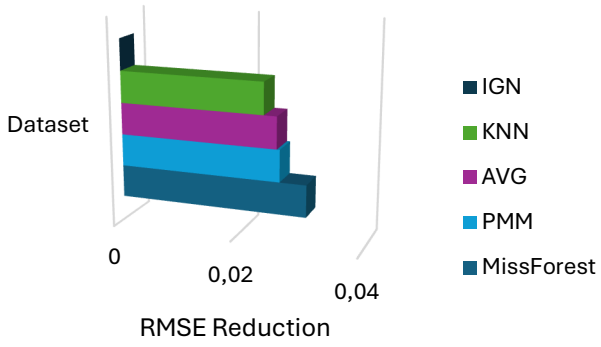

Fig. 2 Reduction in RMSE in each imputed dataset from experiment with respect to IGN

Specifically, the dataset imputed using MF exhibits the most favorable RMSE outcome, closely followed by PMM, with a marginal difference of approximately 0.0295 compared to IGN. PMM and AVG also demonstrate slightly superior RMSE performance relative to KNN. In general, datasets that underwent IGN imputation exhibit RMSE differences exceeding 0.02 when compared to the other imputation methods.

Table 4 and Fig. 3 offer compelling evidence of a substantial improvement exceeding 10% in prediction accuracy when datasets undergo imputation via KNN, PMM, MF, and AVG methods as compared to IGN. Notably, the most notable enhancement in prediction accuracy is achieved when employing the KNN imputation technique. Despite MF having a lower RMSE, it surpasses both AVG and IGN in

terms of prediction accuracy, underscoring its effectiveness in enhancing model performance.

TABLE IV
PREDICTION ACCURACY BETWEEN IMPUTATION METHODS

Dataset	Prediction Accuracy (%)
KNN	48.2
PMM	47.2
MF	46.1
AVG	45.1
IGN	32.5

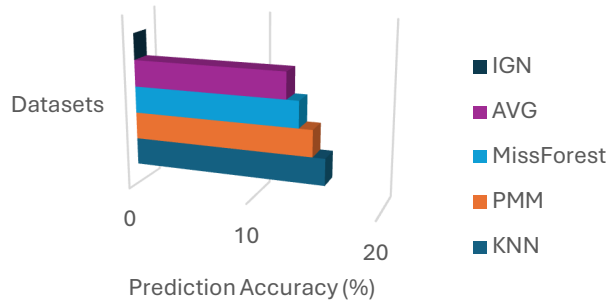

Fig. 3 Improvement in prediction accuracy percentage in each imputed datasets concerning IGN

Table 5 and Fig. 4 comprehensively portray the prediction model's accuracy and sensitivity across each class, gauged through the F-Measure metric, while considering diverse imputation methods applied to the training dataset. This analytical approach offers valuable insights into the model's performance differentials concerning the imputation techniques.

TABLE V
F-MEASURE BETWEEN IMPUTATION METHODS

Dataset	F-Measure		
	UP	NOCHG	DOWN
KNN	0.311	0.220	0.632
PMM	0.425	0.133	0.618
MF	0.489	0.108	0.485
AVG	0.446	0.095	0.581
IGN	0.267	0.282	0.408

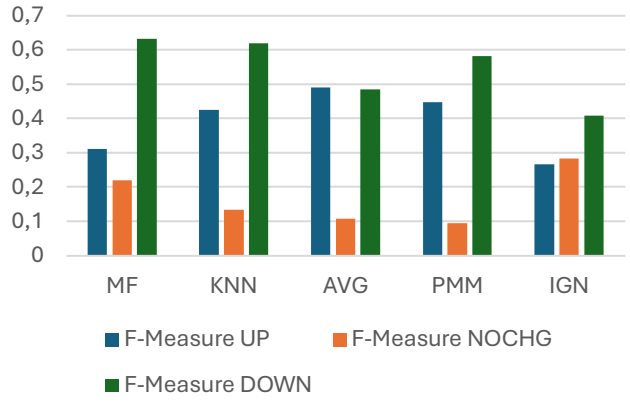

Fig. 4 Comparisons of F-Measure between imputation methods from experiment

Except for IGN, an overarching observation is that both UP and DOWN classes consistently exhibit higher F-Measure values in all datasets than the NOCHG class. This observation aligns with the conceptual proximity between NOCHG and UP/DOWN classes, resulting in a comparatively lower F-Measure for NOCHG during the RCSE prediction process. Consequently, this research emphasizes imputation methods capable of augmenting the F-Measure for the NOCHG class while maintaining the observed F-Measure trends.

Except for IGN, the MF imputation method stands out, showcasing a notably higher F-Measure exceeding 0.2 for the NOCHG class. In contrast, KNN and AVG display F-Measure values for the NOCHG class ranging from 0.1 to 0.5. However, PMM records an F-Measure below 0.1 for the NOCHG class, suggesting that the prediction model's accuracy and sensitivity towards the NOCHG class benefit more significantly from PMM-imputed data during the prediction process.

IV. CONCLUSIONS

This paper successfully presents a comprehensive evaluation of prevalent missing data imputation methods, offering valuable insights into their effectiveness. During the evaluation phase, missing values were subjected to imputation using a range of methods, including AVG, IGN, PMM, KNN, and MF, resulting in the creation of ten distinct datasets. These imputed datasets were subsequently leveraged to train an ANN-based prediction model that forecasts RCSE class values within the testing financial dataset. The prediction model's performance was rigorously assessed using three essential metrics: RMSE, prediction accuracy, and F-Measure.

The evaluation findings position the dataset derived from MF imputation as the leading performer among all the imputed datasets. This dataset consistently outperformed the others across various performance indicators. While PMM and KNN exhibited comparable performance, they surpassed AVG and IGN in multiple facets, underscoring their efficacy in enhancing the overall model outcomes. The IGN dataset consistently exhibits the poorest performance across various metrics, including RMSE, prediction accuracy, and F-measure. Within the IGN dataset, the training dataset size is notably reduced to 1743 instances, with 230 cases removed due to missing value occurrences. This decrease in training size naturally results in a decline in the performance metrics, aligning with similar observations made in prior studies.

Compared to IGN, datasets subjected to MF and PMM imputation methods demonstrate superior performance in both RMSE and prediction accuracy. However, it's worth noting that PMM exhibits a significant decline in the F-Measure for the NOCHG class. Conversely, KNN, which attains the highest prediction accuracy, does not fare well regarding RMSE and F-Measure. Notably, KNN's performance, while outperforming AVG, exhibits only a marginal RMSE difference of 0.0021.

It is essential to acknowledge a limitation of this study, which lies in its exclusive focus on imputing missing values in numerical datasets. Consequently, the outcomes and insights derived from this research primarily apply to datasets with numeric attributes. Future research endeavors may explore extending these imputation methods to datasets with different data types, ensuring a more comprehensive understanding of missing value imputation across diverse data domains.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the International Matching Grant with Project ID UIC241510 from the Universiti Malaysia Pahang Al-Sultan Abdullah (RDU242708). This support is gratefully acknowledged.

REFERENCES

[1] M. S. Gangadhar, K. V. S. Sai, S. H. S. Kumar, K. A. Kumar, M. Kavitha, and S. S. Aravinth, "Machine Learning and Deep Learning Techniques on Accurate Risk Prediction of Coronary Heart Disease," in 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, Feb. 2023, pp. 227–232. doi:10.1109/ICCMC56507.2023.10083756.

[2] X. Kong, W. Zhou, G. Shen, W. Zhang, N. Liu, and Y. Yang, "Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data," vol. 261, p. 110188, 2023, doi:10.1016/j.knosys.2022.110188.

[3] E. Getzen, L. Ungar, D. Mowery, X. Jiang, and Q. Long, "Mining for equitable health: Assessing the impact of missing data in electronic health records," J Biomed Inform, vol. 139, p. 104269, Mar. 2023, doi:10.1016/J.JBI.2022.104269.

[4] K. Psychogyios, L. Ilias, C. Ntanos, and D. Askounis, "Missing Value Imputation Methods for Electronic Health Records," IEEE Access, vol. 11, pp. 21562–21574, 2023, doi: 10.1109/ACCESS.2023.3251919.

[5] B. Agbo, H. Al-Aqrabi, T. Alsboui, M. Hussain, and R. Hill, "Imputation of Missing Clinical Covariates for Downstream Classification Problems," IEEE Access, vol. 11, pp. 102935–102943, 2023, doi: 10.1109/ACCESS.2023.3317775.

[6] P. Buczak, J. J. Chen, and M. Pauly, "Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms," Entropy 2023, Vol. 25, Page 521, vol. 25, no. 3, p. 521, Mar. 2023, doi: 10.3390/E25030521.

[7] G. Shen, W. Zhou, W. Zhang, N. Liu, Z. Liu, and X. Kong, "Bidirectional spatial–temporal traffic data imputation via graph attention recurrent neural network," Neurocomputing, vol. 531, pp. 151–162, Apr. 2023, doi: 10.1016/J.NEUCOM.2023.02.017.

[8] L. Li, Y. Wang, H. Wang, S. Hu, and T. Wei, "An Efficient Architecture for Imputing Distributed Data Sets of IoT Networks," IEEE Internet Things J, vol. 10, no. 17, pp. 15100–15114, Sep. 2023, doi: 10.1109/JIOT.2023.3264609.

[9] G. Batista and M.-C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method," in Hybrid Intelligent Systems, ser Front Artificial Intelligence Applications, Jan. 2002, pp. 251–260.

[10] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," Journal of Systems and Software, vol. 85, no. 11, pp. 2541–2552, Nov. 2012, doi: 10.1016/J.JSS.2012.05.073.

[11] Y. He and D. Pi, "Improving KNN Method Based on Reduced Relational Grade for Microarray Missing Values Imputation," IAENG Int J Comput Sci, vol. 43, no. 3, pp. 356–362, 2016.

[12] J.-H. Hsu, C.-H. Wu, W.-K. Wang, H.-Y. Su, E. C.-L. Lin, and P. S. Chen, "Digital Phenotyping-Based Bipolar Disorder Assessment Using Multiple Correlation Data Imputation and Lasso-MLP," IEEE Trans Affect Comput, pp. 1–14, 2023, doi10.1109/TAFFC.2023.3299607.

[13] I. D. Irawati, A. B. Suksmono, I. J. M.Edward, "An Interpolation Comparative Analysis for Missing Internet Traffic Data," Proceedings of the 3rd International Conference on Electronics, Communications and Control Engineering, pp. 26-30, 2020, doi:10.1145/3396730.3396740

[14] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[15] A. K. Waljee et al., "Comparison of imputation methods for missing laboratory data in medicine," BMJ Open, vol. 3, no. 8, p. e002847, Aug. 2013, doi: 10.1136/bmjopen-2013-002847.

[16] J. You, J. L. Ellis, S. Adams, M. Sahar, M. Jacobs, and D. Tulpan, "Comparison of imputation methods for missing production data of dairy cattle," animal, p. 100921, Jul. 2023, doi:10.1016/j.animal.2023.100921.

[17] B. Gong, Z. Xu, C. Lin, and D. Wu, "Heterogeneous Traffic Flow Detection Using CAV-Based Sensor With I-GAIN," IEEE Access, vol. 11, pp. 32616–32627, 2023, doi: 10.1109/ACCESS.2023.3263720.

[18] G. Vink, L. E. Frank, J. Pannekoek, and S. van Buuren, "Predictive mean matching imputation of semicontinuous variables," Stat Neerl, vol. 68, no. 1, pp. 61–90, Feb. 2014, doi: 10.1111/stan.12023.

[19] J. Du and L. Zhou, "Improving financial data quality using ontologies," Decis Support Syst, vol. 54, no. 1, pp. 76–86, Dec. 2012, doi: 10.1016/j.dss.2012.04.016.

[20] Idris NF, Ismail MA, Jaya MIM, Ibrahim AO, Abulfaraj AW, Binzagr F (2024) Stacking with Recursive Feature Elimination-Isolation Forest for classification of diabetes mellitus. PLoS ONE 19(5): e0302595. https://doi.org/10.1371/journal.pone.0302595.