Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?

Youran Zhou(✉), Mohamed Reda Bouadjenek, and Sunil Aryal

School of Information Technology, Deakin University, Geelong, VIC, Australia
{echo.zhou,reda.bouadjenek,sunil.aryal}@deakin.edu.au

Abstract. Missing data poses a significant challenge in real-world data analysis, prompting the development of various imputation methods. However, existing literature often overlooks two critical limitations. Firstly, many methods assume a Missing Completely At Random (MCAR) mechanism, which is relatively easy to handle but may not reflect real-world scenarios where data is often missing due to some underlying mechanisms (issues/problems) that are often unknown. This type of missing data is categorized as Missing At Random (MAR) and Missing Not At Random (MNAR). Secondly, the effectiveness of these methods is primarily assessed solely in terms of imputation accuracy using metrics such as Root Mean Square Error (RMSE), ignoring the practical utility of imputed data in downstream tasks. In this study, we comprehensively compare a broad spectrum of missing data imputation techniques, ranging from traditional statistical methods to advanced machine and deep learning approaches. Our evaluation considers their effectiveness in handling various missing mechanisms across different missing parameters. Furthermore, we assess the imputed data's quality not only in terms of RMSE but also its impact on downstream tasks, such as classification, regression, and clustering. Contrary to common assumptions, our findings reveal that the superiority of complex deep learning-based methods is not guaranteed over simple traditional techniques. Moreover, relying solely on RMSE for evaluation can be misleading. Instead, selecting an imputation method should prioritise its effectiveness in enhancing the performance of learning algorithms in downstream tasks.

Keywords: Data imputation · Missing Data · Missing Mechanism · MCAR · MNAR · Tabular data · Deep learning · Machine learning

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-70381-2_7.

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14950, pp. 100–115, 2024.
https://doi.org/10.1007/978-3-031-70381-2_7

1 Introduction

Missing data refers to the loss of values or information within fields or attributes in a dataset. This phenomenon occurs for various reasons during data collection, storage, or processing. In real-world applications, the precision of machine learning models and statistical analyses hinges significantly on the quality of the data used. Missing data will decrease its quality and pose significant challenges in pattern mining. Mishandling these missing values can introduce biases, compromise the generalizability of findings, and impede the development of robust models. Several missing data imputation methods have been proposed in the literature. They range from simple traditional Statistical-based to Machine learning-based (ML), and Deep learning-based (DL) methods.

It is crucial not only to develop suitable missing data imputation methods but also to establish appropriate evaluation metrics to assess their effectiveness. Commonly used evaluation metrics include Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), which aim to quantify the discrepancy between actual and imputed values. However, solely relying on metrics like RMSE might underestimate the true impact of the imputed data. RMSE and MAE fail to capture the underlying variability or distribution, potentially leading to poor results in downstream tasks. Moreover, while RMSE metrics provide a direct measure of data quality, they may not be applicable in real-world scenarios where true value is not available to compare, rendering such quantitative assessment unfeasible. Therefore, it is more prudent to evaluate imputation techniques based on their effectiveness in downstream tasks.

Existing literature on missing data imputation often overlooks two critical limitations. Firstly, the majority of methods assume the occurrence of Missing Completely At Random (MCAR) mechanisms. This scenario is relatively easy to handle as the distribution of data is not distorted significantly. However, this assumption is mostly violated in real-world applications, where data is often missed because of some underlying issues or causes. i.e., the missing mechanism is not completely random. Missing mechanisms like Missing Not At Random (MNAR) and Missing At Random (MAR) introduce more complex assumptions and possess limited prior knowledge regarding the distribution of data. Secondly, the evaluation metrics for imputation accuracy predominantly rely on quantitative measures such as RMSE, thus disregarding the comprehensive practical utility of imputed data in downstream tasks.

In this study, we focus on tabular data for various missing mechanisms and conduct a systematic experiment involving three types of imputation methods mentioned above. Our primary focus is on conducting comprehensive evaluations of imputation methods in terms of their ability to handle different missing mechanisms and missing parameters and their effectiveness in downstream tasks, including regression, classification, and clustering. Furthermore, we offer insights into future directions for refining the evaluation metrics of the data imputation problem.

2 Related Work

Little and Rubin [13] underscore the importance of robust statistical strategies for addressing missing data effectively, highlighting the necessity of imputation techniques. A lot of studies have explored missing data imputation methods, typically classified into three primary categories: Statistical-based, ML-based, and DL-based approaches. Statistical-based methods involve estimating and replacing missing values using statistical principles. ML-based approaches leverage unsupervised or supervised learning to predict missing values, leveraging available non-missing data information [17,25,27]. Recent efforts have also focused on DL-based imputation methods [8,18,28,29]. Previous comparison and survey studies have extensively explored and summarized existing methodologies, shedding light on their relative strengths and limitations. In the work by Lin et al. [12], over one hundred of articles were summarized along with their study. They found that only a small portion of articles mention using their imputation methods for classification downstream tasks, with hardly any articles involving both quantitative and downstream task evaluation. Harel et al. [6] provide a summary of multiple imputation techniques, evaluating the quality of imputed data using estimates of sensitivity, standard error, lower and upper confidence intervals, and other statistical properties. Articles such as [15] focus on quantitative analysis, while articles like [9] also consider downstream tasks but utilize only a limited number of imputers. Alabadla et al. [1] conducted a survey on ML-based imputations, noting the scarcity of articles working on all three missing mechanisms: MCAR, MAR, and MNAR. Most articles in this survey use RMSE and accuracy as evaluation metrics, often combined with classification downstream tasks. Miao et al. [19] provide a comprehensive experimental survey. However, they only paid minimal attention to classification tasks as the post-imputation task. Studies such as [4,14] focus on time series and imputation in healthcare but lack comprehensive comparative studies and do not mention the evaluation process.

From our study of the existing literature, we found that there is a lack of a comprehensive experimental survey study on the utility of different data imputation methods using various evaluation metrics in supervised and unsupervised learning downstream tasks. Additionally, imputation algorithms vary in problem settings, missing data characteristics, and data selection. Therefore, in this paper, we conduct a systematic and comprehensive experimental study of statistical, machine learning, and deep learning types of missing data imputation under all three types of missing mechanisms and missing rates, using quantitative and downstream tasks to fully evaluate the imputation methods.

3 Background

We define a complete data matrix with k variables and n instances, $\boldsymbol{X} = (x_1, \ldots, x_n)^T \in \mathcal{X}^k$. In the context of missing data, each sample can be divided into the observed part X^o and the missing part X^m, where X^o contains no

missing values and X^m contains the missing values. Missing Parameter Ψ represents the process of generating missing data, including the missing mechanism, missing rate, and all factors that impact this process. Rubin [23] proposed that missing data could be catagorized into three groups: MCAR, MAR and MNAR.

MCAR mechanism implies that the missingness is unrelated to the specific observation under study or any other variables within the dataset. Data values are missed completely randomly. **MAR** mechanism suggests the missingness of data can be predicted based on other variables within the study, but not directly from the missing data itself. **MNAR** exhibit missingness that is directly linked to the value of the missing observation itself X^m. It is important to recognize that MNAR poses a more challenging scenario, as the missing data cannot be ignored or dropped without introducing bias into subsequent analyses.

Table 1 provides examples of each missing mechanism using a dataset with a fully observed variable **IQ** and a variable **R** (*job performance ratings*) with missing entries denoted by '?'. For MCAR, missingness is independent of any data. For MAR, missingness in **R** depends on **IQ** scores below 90. For MNAR, missingness in **R** occurs when ratings are below 10.

Table 1. An Example of different missing mechanisms proposed by [3]

IQ	$\mathbf{R}_{Full}$	$\mathbf{R}_{MCAR}$	$\mathbf{R}_{MAR}$	$\mathbf{R}_{MNAR}$
78	9	?	?	?
84	10	?	?	10
87	7	7	?	?
92	9	9	9	?
94	11	11	11	11
96	7	?	7	?
105	10	10	10	10
106	15	15	15	15
112	10	?	10	10
115	14	14	14	14
134	12	?	12	12

4 Imputation Methods

For our discussion, we categorize imputation algorithms into three groups: Statistical methods, ML methods, and DL methods. Although DL is a subset of ML, we separate it to highlight its unique features and recent advancements. These categories reflect different levels of model complexity. In this experiment, we selected the state-of-the-art and most representative imputers from each category.

4.1 Statistical Methods

Statistical imputation estimates and imputes missing values using statistical properties, often serving as baseline methods. **Zero (ZR)** Imputer assigns a value of 0 to each missing value, acting as a placeholder. **Mean**, median, and mode imputation calculate and use the mean, median, or mode of non-missing values to impute missing data [24]. **Random Imputer (RD)** replaces missing values with random values between the observed minimum and maximum.

4.2 Machine Learning Methods

Most machine learning imputation methods treat imputation as a supervised learning task, predicting missing values using observed data. The **K-Nearest**

Neighbor (KNN) [11] model imputes missing values by using the values from the nearest neighbors based on a chosen distance function. **Matrix Factorization (MF)** [22] decomposes the dataset into lower-dimensional matrices to reconstruct missing entries. **Multiple Imputation by Chained Equations (MICE)** [27] uses regression models to iteratively impute missing values for each variable based on other observed variables. **eXtreme Gradient Boosting (XGB)** [17] and **MissForest (MisF)** [25] are tree-based imputation methods. MissForest is suitable for mixed-type missing data and starts by filling missing values with mean imputation. It then iteratively trains a random forest on the observed data to update the imputation until convergence. Similarly, XGB uses XGBoost as the predictor to iteratively update missing values. **Optimal Transport (OT)** [20] and **HyperImputer (HI)** [10] enhance existing imputation methods. OT uses optimal transport to define a loss function, improving Multiple Imputation by MICE and Multi-layer Perceptron. HyperImpute is a flexible iterative imputation framework that automatically configures column-wise models and their hyperparameters, with effectiveness depending on the chosen base imputer.

These ML methods do not incorporate specialized components to address MAR and MNAR assumptions. As a result, these methods are not expected to be robust enough to handle all missing data mechanisms effectively.

4.3 Deep Learning Methods

Recently, DL-based imputation methods have gained significant attention. Models such as Generative Adversarial Networks (GAN) and Variational Autoencoder (VAE) have emerged as prominent tools for addressing missing data imputation [8,18,28,29]. **Generative Adversarial Imputation Nets (GAIN)** [28] is a GAN-based imputer that uses a generator to fill in missing data based on observed components. A discriminator distinguishes between observed and imputed data, guided by a hint vector that indicates the missing pattern. VAEs consist of an encoder that maps input data to a latent space distribution and a decoder that generates data samples from this distribution. For imputation tasks, VAE imputers use an encoder to map incomplete data and a mask matrix to a latent space, and a decoder to generate imputed results from this distribution [8,16,18,21]. **Missing Importance-Weighted AutoEncoder (Mi)** [18] and **Not-Missing Importance-Weighted AutoEncoder (NMi)** [8] enhance VAEs by adding information during training to handle MAR and MNAR data effectively. Diffusion models [7] draw inspiration from non-equilibrium thermodynamics, using a Markov chain of diffusion steps to introduce random noise to data gradually. They then learn to reverse this process to generate desired data samples from the noise. **Table Conditional Score-based Diffusion Models (CSDI)** [26,29] employ conditional score-based diffusion, allowing them to process and impute incomplete datasets effectively.

These methods represent three typical DL frameworks for missing data imputation. Note that DL-based imputation methods exhibit stochastic behavior due to the introduction of random noise in their generative processes. This stochastic

nature means that outputs may vary with each run, even with identical inputs, although careful control of the model's random processes can mitigate this variability.

5 Experiments

We conducted comprehensive experiments to investigate the effectiveness of existing missing value imputation methods in handling different missing mechanisms (MCAR, MAR, and MNAR) and missing parameters. We assess the quality of imputed data using quantitative metrics like RMSE and its utility in downstream tasks such as classification, clustering, and regression. The code to reproduce these experiments is available at https://github.com/echoid/ML_DL_Missing_Data_Imputation. Detailed implementation and complete results are provided in the supplementary material.

5.1 Datasets and Experimental Setting

Dataset. We utilized ten datasets from the UCI Machine Learning Repository[1], as summarized in Table 2. These datasets are frequently used in prior studies on missing value imputation [19,20]. Our experiment focuses on purely numerical data, primarily continuous variables, since not all imputation methods can handle categorical variables.

Table 2. UCI Data Summary. Task indicates the downstream task associated with datasets (C-Classification, R-Regression. Classification datasets can also be used for clustering)

Dataset	Bank	Cali	Climate	Concre	Qsar	Red	Sonar	White	Yachts	Yeast
#Inst	1372	20640	540	1030	1055	1500	208	4898	308	1484
#Dim	5	9	20	8	41	11	60	11	6	8
Task	C	R	C	C	C	R	C	R	R	C

Missing Data Generation. For experimentation purposes, missing data is created from the complete dataset. Our experiments cover MCAR, MAR and MNAR scenarios. Given the absence of well-established standards for generating MNAR data in the existing literature, we adapted implementations of MNAR generation as discussed in [5,8,20]. Each missing method is characterized by its own missing parameter Ψ. In this experiment, we use Ψ to approximate the missing rate for most cases (except MNAR-P). Missing rates of 30%, 50%, and 70% are used to indicate varying severities of missing data.

[1] https://archive.ics.uci.edu/.

- **MCAR**: This method randomly selects data to be missing. We use Ψ values of 0.3, 0.5, and 0.7 to represent missing rates of 30%, 50%, and 70%, indicating slight, partial, and severe missingness across all features.
- **MAR**: We utilize an implementation from the OT [20] with missing parameters Ψ set at 0.3, 0.5, and 0.7.
- **MNAR-P**ercentile: Inspired by the NMi paper [8], this method involves dividing each column into quantile ranges (Q1, Q2, Q3, Q4) and selecting specific blocks as missing. We designate Q1 & Q4, Q2 & Q3, and Q2 & Q4 to represent missing values at the distribution's tails, within the central range, and across two non-adjacent segments, respectively. The missing rate is set at 50%.
- **MNAR-L**ogistic: This method uses the OT implementation [20] with a default proportion of variables (p) set to 0.3. The missing rate parameter Ψ is set to 0.3, 0.5, and 0.7.
- **MNAR-D**iffuse: This method, representing a diffuse MNAR approach, is refined from [5]. We designate 50% of the columns as missing and the remainder as observed, with the missing rate parameter Ψ set to 0.3, 0.5, and 0.7.

Table 3. Imputation Methods Summary

Model Name	Type	Subtype
RD, ZR, Mean	Statistical	Baseline
KNN, MF[a]	ML	-
MICE	ML	Regression-based
XGB[b], MisF	ML	Tree-based
OT, HI	ML	Enhance ML Model
GAIN	DL	GAN-based
Mi, NMi	DL	VAE-based
CSDI	DL	Diffusion-based

[a] https://pypi.org/project/fancyimpute/
[b] https://github.com/sjtupig/MissingImputer

Imputation Methods. We deploy a diverse array of imputation techniques in our experiments. Table 3 summrized the imputation methods we use in our experiments. The implementation code and parameter configurations for ZR, Mean, KNN, MisF, and MICE are sourced from the *Sci-Kit Learning* Package with default parameter settings. The remaining methods are obtained from their respective implementation repositories, also using default parameter settings.

Experimental Setup. First, for each dataset, we apply column-wise min-max scaling to ensure all values fall within the range of 0 to 1. This helps with anomaly detection, as imputed samples should reside within this range, allowing us to observe if the model imputations fall outside this range. Next, we conduct

a five-fold data split. For each fold, we partition 10% of the training set as the validation set. Each imputer is trained five times using different training and validation sets, iterating through the data folds. Finally, imputation is performed on both the test and training sets, preparing them for evaluation.

Evaluation Metrics. We adopt a multi-faceted approach to assess the performance of imputation methods, incorporating both quantitative metrics and machine learning utility. The **RMSE** quantifies the discrepancy between imputed and actual values, with lower RMSE values indicating better results. However, relying solely on RMSE may not capture the full utility of the imputed data. Therefore, we also use the average **Pearson correlation** between imputed and actual values for each column with missing data. Pearson correlation measures the strength of the linear relationship between two variables, with values ranging from -1 to 1, where a value closer to 1 indicates better results. Our goal is to achieve a strong linear relationship with a correlation coefficient close to 1 between imputed and actual values.

Additionally, we assess the utility of imputed datasets for downstream tasks. A good imputation method should enable the imputed dataset to perform similarly to the complete data in these tasks. We evaluate both supervised and unsupervised learning tasks. In supervised learning experiments, we train machine learning models on the imputed training set and evaluate their performance using complete test sets. For classification tasks, we use Logistic Regression, MLP classifier, and SVM classifier, evaluating performance with metrics such as the **F1** score. For regression tasks, we use Ridge Regression, MLP regressor, and SVM regressor, recording **RMSE** values. The selection of three models for regression and classification helps to avoid bias. In unsupervised learning, we perform clustering using DBSCAN on both complete and imputed datasets, aiming to observe similarities in clustering labels. We compare results using Adjusted Mutual Information (**AMI**). Generally, higher evaluation metrics indicate better results, except for RMSE in regression tasks. All models are implemented using *scikit-learn* with default parameter settings.

5.2 Experimental Results

We systematically compared the imputation results of different methods across various missing mechanisms, using different missing rates and datasets. Due to page limits, we present our key summarised results and findings of our experiments in this section.

Ranking Analysis. To summarize the results, we ranked the performance of imputation methods from best to worst for each dataset, with the best result assigned the lowest rank. We then calculated the average rank across different datasets. Figure 1 shows the average rankings of various imputation methods across five missing mechanisms, evaluated by RMSE, correlation coefficient between imputed and true values, and performance in three downstream tasks-SVM Classification, DBSCAN Clustering, and SVM Regression. These results

correspond to a missing rate of 0.5. For MNAR-P, the missing data is presented using blocks Q1 and Q4.

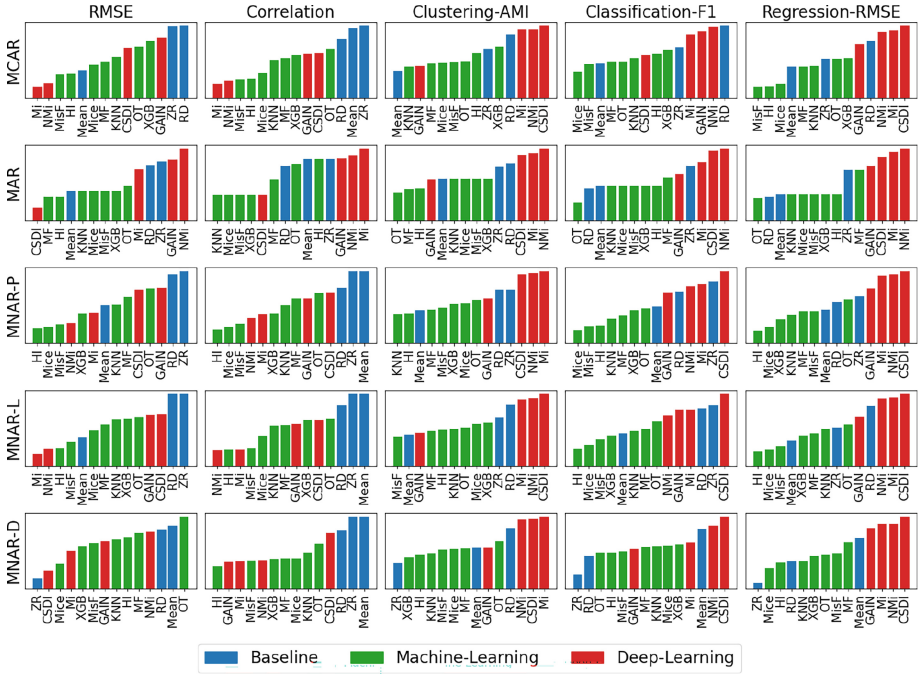

Fig. 1. Average ranking of different imputation methods w.r.t. RMSE and correlation coefficient of imputed and true value (first two columns) along with performances in three downstream tasks (last three columns) for five different missing mechanisms. The lower the rank, the better the imputation method.

Regarding **RMSE** of imputed values against the true values (First Column in Fig. 1):

- In the MCAR missing mechanism, among all datasets, Mi and NMi, two VAE-based DL methods lead the overall performance, followed by the two ML methods of MisF and Hi. Surprisingly, the simple statistical baseline of mean imputation yields acceptable results (top 5 out of 14 methods compared).
- For MAR, the CSDI model emerges as the best performer, followed by MF, HI, and Mean. Once again, mean imputation demonstrates commendable results. DL-based methods exhibit subpar performance, GAIN and NMi are even worse than the simple baselines of ZR and RD imputers. Additionally, despite being designed specifically for MAR, the MI method fails to produce good results, indicating its lack of robustness across all MAR scenarios.
- MNAR-P and MNAR-L follow a similar trend to MCAR, with machine learning models leading the results. Conversely, MNAR-D exhibits different performance, with the ZR imputer leading, followed by DL and ML methods in

the top 5. The superior performance of the zero imputer can be attributed to the nature of Diffuse MNAR and dataset properties, where many values close to zero are missing. Notably, NMi is designed specifically for MNAR data, did not consistently outperform across all MNAR scenarios, indicating its limited robustness.

Overall, by examining the RMSE, we conclude that deep learning methods are adept at handling missing data across various missing mechanisms.

Regarding **Correlation coefficient** of imputed and true values (Second Column in Fig. 1):

- The correlation values for RD, ZR, and Mean are expected to be the lowest, as these methods do not involve learning and rely on random number generation or single-value replacement, lacking statistical significance. Generally, the trends for MCAR and MNAR-P are similar to those for RMSE values. For MAR, most deep learning methods, except for CSDI, performed worse than the baseline methods. Across other missing mechanisms, the two VAE-based methods, MNi and Mi, consistently perform well, while the performance of CSDI and GAIN fluctuates. ML-based imputation methods, HI and MisF, generally exhibit robust performance across all missing mechanisms.

Overall, the correlation analysis does not entirely align with the conclusions drawn from RMSE. While DL-based methods showed promising performance in correlation analysis, they did not consistently outperform other methods.

Regarding **Downstream tasks** (Columns 3 for Clustering, 4 for Classification, and 5 for Regression in Fig. 1):

- Interestingly, the classification and regression results exhibit significant differences from the previously discussed RMSE results. Across all missing mechanisms, ML-based methods lead the results, while DL-based methods consistently underperform compared to ML-based methods, none are ranked within the top 5.
- In the MNAR-D scenario, the ZR imputer leads the results, highlighting its suitability for this specific condition.
- Overall, traditional ML methods outperform others in classification and regression tasks across most missing mechanisms. Conversely, DL methods consistently show inferior performance in these tasks across all scenarios and datasets.
- For clustering results, clear trends in model performance are not evident, but DL methods consistently underperform in clustering tasks as well.

In summary, these plots reveal a clear trend: while DL methods may show promise in terms of RMSE and correlation, they struggle with downstream tasks. Conversely, machine learning-based methods consistently demonstrate robust performance across all types of missing data. This conclusion is supported across different classifiers and regressors (see y Material).

Missing Rate Analysis. Table 4 summarizes the average RMSE values across 10 datasets with varying missing rates. It presents the average RMSE values

for different missing rates under MCAR, MAR, and MNAR-L. MNAR-D and MNAR-P are excluded because MNAR-P has different missing parameters, and MNAR-D's missing ratio and parameters are not linearly independent. Overall, we observe that as the missing rate increases, the RMSE value also increases, indicating that models receive less information with higher missing rates, leading to poorer performance. MCAR appears to be the easiest scenario to handle, with generally smaller RMSE values compared to other missing mechanisms in most cases.

Table 4. Average RMSE at different missing parameters/rates. RSME values are scaled by a factor of 10 to show the differences up to three decimal places

Ψ	Baseline			Machine Learning							Deep Learning			
	RD	ZR	Mean	KNN	MF	Mice	MisF	XGB	OT	HI	GAIN	Mi	NMi	CSDI
MCAR														
0.3	4.30	4.22	1.86	1.86	1.96	1.59	**1.60**	4.16	2.16	1.66	2.34	1.62	**1.60**	2.22
0.5	4.29	4.22	1.85	2.02	1.99	2.31	1.80	3.98	2.20	1.83	2.64	**1.69**	1.70	2.24
0.7	4.29	4.23	1.86	2.09	2.02	3.08	2.04	3.39	2.20	2.06	3.12	**1.81**	1.92	2.28
MAR														
0.3	4.27	4.05	2.41	2.41	2.21	2.41	2.41	2.41	2.60	2.36	4.08	3.17	74.30	**2.13**
0.5	4.25	4.10	2.53	2.53	2.38	2.53	2.53	2.53	2.59	2.35	4.08	3.31	78.99	**2.11**
0.7	4.39	3.95	2.76	2.76	2.44	2.76	2.76	2.76	2.70	2.48	4.12	3.12	81.68	**2.40**
MNAR-L														
0.3	4.27	4.23	1.88	1.97	1.98	1.99	1.62	3.60	2.15	**1.60**	2.23	1.63	1.64	2.24
0.5	4.28	4.23	1.88	2.06	2.01	2.15	1.82	2.59	2.16	1.78	2.44	**1.69**	1.70	2.22
0.7	4.28	4.20	1.90	2.12	2.04	3.24	2.05	2.59	2.19	2.01	2.89	**1.81**	1.86	3.04

6 Discussion

We aim to discuss why DL methods yield inconsistent results between RMSE and downstream tasks. Using Permutation Feature Importance [2], we measure each feature's contribution to a model's performance on the Banknote dataset, a binary classification problem with four features and an MCAR missing mechanism with a 50% missing rate. We employed SVM as the classifier due to the dataset's simplicity, which aids in result visualization. We consider four scenarios: the complete dataset and data imputed using OT, NMi, and CSDI methods. An SVM classifier trained on the complete dataset provides an upper-bound guideline for F1 and accuracy scores.

Results in Fig. 2 show that OT exhibits relatively low RMSE and high F1 and accuracy scores, outperforming NMi. CSDI, despite its low RMSE, fails in

the machine learning task. The complete dataset emphasizes feature 1, with OT closely aligning with this model. For NMi, features 1 and 3 are significant.

Scatter plots (Fig. 3) comparing actual to imputed values reveal further insights. The plots range from 0 to 1 due to min-max normalization, with RMSE values indicating imputation accuracy. Complete data shows perfect alignment with zero RMSE. Among imputation methods, NMi has the lowest RMSE. However, OT achieves the best classification F1 score, suggesting that lower RMSE does not always correlate with better classification. CSDI shows higher RMSE and significant scatter, reflecting lower accuracy and deviations from actual values, impacting downstream tasks. Additionally, CSDI's scatter plots exhibit a wider range and larger dispersion, indicating less precise imputation. Performance beyond this range is uncertain, necessitating further analysis.

This analysis suggests that low RMSE imputation methods do not necessarily preserve crucial information for downstream tasks.

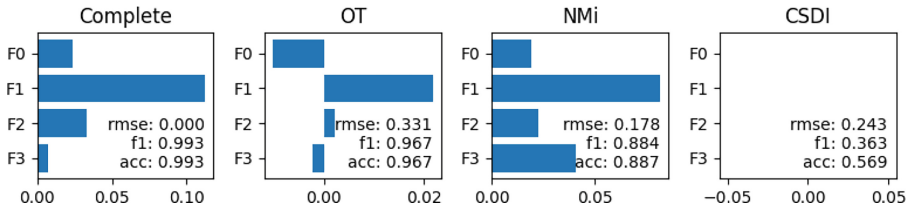

Fig. 2. Permutation feature importance analysis on Banknote Data under the MCAR missing mechanism with 0.5 Missing Rate. Results for the CSDI method are absent due to failure, i.e., features are not ranked.

To further evaluate the performance of the CSDI imputer, we utilized the confusion matrix (See Fig. 4) and discovered that the model consistently predicts the same outcome (Negative) for all instances. This suggests that the model struggles to distinguish between the decision boundaries for different outcomes. Consequently, we selected feature 1 and feature 2-identified as the two best features from the complete dataset-from the CSDI-imputed data. We used these features to retrain the SVM model and visualize the resulting decision boundary in Fig. 5. Within this decision boundary plot, we depicted the training set (imputed data shown on blue squares and yellow triangles) and the test data (blue circles) to evaluate the alignment of the decision boundary with the test data.

Upon analyzing Fig. 5, we noticed that both the x-axis and y-axis have been scaled to large values due to some imputed data points approaching these excessively large values. This is abnormal, as all data was scaled using min-max normalization and should fall within the range of 0 to 1. It appears that some CSDI imputed values failed to generate data within the valid range. However, the test points, marked with blue circles, are positioned in the middle of the plot as expected, within the 0 to 1 range. Further examination revealed that

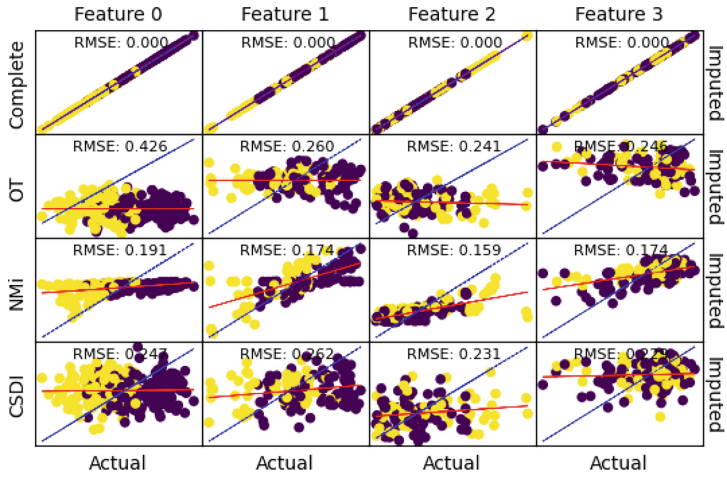

Fig. 3. Scatter plots comparing actual to imputed values for Banknote dataset features under MCAR (0.5 missing rate) using four methods. Red lines are best fit lines expected to align with the diagonal. RMSE values indicate imputation accuracy. Red and purple indicate the two labels (Color figure online).

the majority of the data points are classified as blue, yielding only four distinct regions. However, the plot's expansive scale causes the yellow region to stretch well beyond the 0 to 1 range, making classification into the yellow group virtually impossible for the test data. This large scale suggests that certain data points fall outside the 0 to 1 range. These extreme values, generated by CSDI, disrupt our data distribution and compromise the SVM decision boundary's accuracy. Although these outliers are limited in number and can go unnoticed within a large dataset when computing RMSE, they artificially lower the RMSE by being averaged down. However, when building machine learning models, these invalid points often have a significant impact on the model's performance.

In summary, only relying on RMSE for evaluation can be misleading. RMSE calculates an average value, and anomalies resulting from failed algorithms or invalid inputs can compromise downstream tasks. Even if the RMSE appears acceptable due to a limited number of invalid inputs, it can still negatively impact downstream tasks. Moreover, if we focus primarily on RMSE, detecting these abnormalities becomes challenging. Therefore, solely depending on RMSE can lead to the overestimation of imputation results.

7 Conclusions and Future Directions

This study presents a comprehensive comparison of various imputation methods across MCAR, MAR, and MNAR missing mechanisms, utilizing qualitative analysis alongside downstream tasks to evaluate data utility. The key findings are summarized as follows:

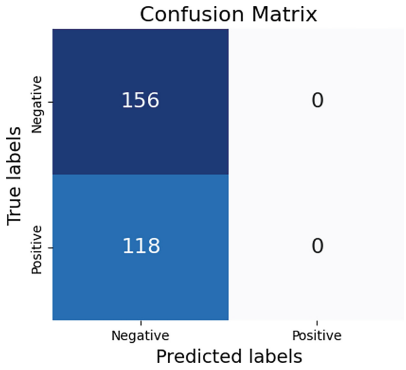

Fig. 4. Confusion Matrix for prediction results using the CSDI imputed training data

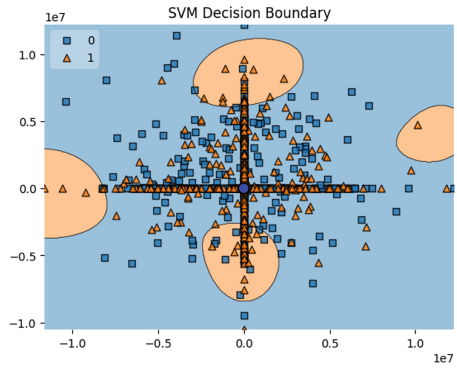

Fig. 5. Decision Boundary for SVM trained using the CSDI imputed training data

- Missing Mechanisms: Imputation methods perform relatively well under MCAR but face challenges with MAR and MNAR due to specific assumptions and complexities, leading to varied performance outcomes.
- Impact of Missing Rate: Higher missing ratios adversely affect imputation method performance, resulting in reduced information available for learning and consequently higher RMSE values.
- Imputation Model Choice: In our experiments, certain statistical methods deliver feasible results, especially in complex missing mechanism cases. Conversely, machine learning-based methods exhibit greater robustness compared to deep learning-based approaches. They excel in quantitative analysis and consistently produce reliable outcomes in downstream task evaluations. While deep learning-based methods show potential in qualitative analysis, they often underperform in downstream tasks. This underperformance could be attributed to the small tabular datasets used in our study, which may not provide sufficient training data for deep learning models to fully demonstrate their capabilities.

Data often goes missing in real-world applications due to various underlying phenomena, such as sensor malfunctions, communication link failures, or certain groups of participants refusing to provide specific information (e.g., age, income, etc.) in surveys. This leads to missing data mechanisms classified as MAR and MNAR, which present significant challenges for imputation methods as they struggle to cope with these cases. Importantly, the true value of imputing missing data is realized only when it is used in subsequent analytical tasks; without these tasks, imputation has little utility. Our study reveals that not every DL imputation method performs adequately in these downstream tasks. This inadequacy often stems from the inability of ML/DL-based methods to impute values within a valid range. Moreover, this shortfall might not be evident when using RMSE as the sole evaluation metric, potentially leading to unnoticed failures

in downstream tasks. Consequently, while these methods may appear effective in quantitative analysis, their utility in practical, downstream tasks is questionable. Therefore, it is suggested that practitioners carefully consider the end-use of imputed data and choose an imputation method that not only addresses the nature of the missing data but also meets the requirements of their specific downstream analytical tasks.

Moving forward, several research challenges and future directions for data manipulation are identified:

- Data Utility: Beyond RMSE, a broader set of metrics is necessary for a true evaluation of imputation quality across analytical tasks.
- Missing Mechanisms Exploration: Existing research largely ignores MAR and MNAR, more common than MCAR, underscoring the need for techniques adept at handling these scenarios.
- Handling Different Data Types: Research must broaden to address missing data in both discrete and categorical forms, moving beyond the current focus on numeric data to tackle real-world challenges effectively.

Addressing these challenges and pursuing these future directions will enhance the effectiveness and applicability of imputation methods in real-world data analysis scenarios.

Acknowledgement. This work is partially supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4003.

References

1. Alabadla, M., et al.: Systematic review of using machine learning in imputing missing values. IEEE Access **10**, 44483–44502 (2022)
2. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
3. Enders, C.K.: Applied Missing Data Analysis. Guilford Publications, New York (2022)
4. Fang, C., Wang, C.: Time series data imputation: a survey on deep learning approaches (2020)
5. Gomer, B., Yuan, K.H.: Subtypes of the missing not at random missing data mechanism. Psychol. Methods **26**(5), 559 (2021)
6. Harel, O., Zhou, X.H.: Multiple imputation: review of theory, implementation and software. Stat. Med. **26**(16), 3057–3077 (2007)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
8. Ipsen, N.B., Mattei, P.A., Frellsen, J.: Not-MIWAE: deep generative modelling with missing not at random data. arXiv preprint arXiv:2006.12871 (2020)
9. Jäger, S., Allhorn, A., Bießmann, F.: A benchmark for data imputation methods. Front. Big Data **4**, 693674 (2021)
10. Jarrett, D., Cebere, B.C., Liu, T., Curth, A., van der Schaar, M.: Hyperimpute: generalized iterative imputation with automatic model selection. In: International Conference on Machine Learning. pp. 9916–9937. PMLR (2022)

11. Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., Tseng, G.C.: Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinform. **15**(1), 1–12 (2014)
12. Lin, W.C., Tsai, C.F.: Missing value imputation: a review and analysis of the literature (2006–2017). Artif. Intell. Rev. **53**, 1487–1509 (2020)
13. Little, R.J., Rubin, D.B.: Bayes and multiple imputation. In: Statistical Analysis with Missing Data, pp. 200–220 (2002)
14. Liu, M., et al.: Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. Artif. Intell. Med. **142**, 102587 (2023). https://doi.org/10.1016/j.artmed.2023.102587. https://www.sciencedirect.com/science/article/pii/S093336572300101X
15. Luo, Y.: Evaluating the state of the art in missing data imputation for clinical data. Briefings in Bioinformatics **23**(1), bbab489 (2021). https://doi.org/10.1093/bib/bbab489
16. Ma, C., Zhang, C.: Identifiable generative models for missing not at random data imputation. In: Advances in Neural Information Processing Systems, vol. 34, pp. 27645–27658 (2021)
17. Madhu, G., Bharadwaj, B.L., Nagachandrika, G., Vardhan, K.S.: A novel algorithm for missing data imputation on machine learning. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 173–177. IEEE (2019)
18. Mattei, P.A., Frellsen, J.: MIWAE: deep generative modelling and imputation of incomplete data sets. In: International Conference on Machine Learning, pp. 4413–4423. PMLR (2019)
19. Miao, X., Wu, Y., Chen, L., Gao, Y., Yin, J.: An experimental survey of missing data imputation algorithms. IEEE Trans. Knowl. Data Eng. (2022)
20. Muzellec, B., Josse, J., Boyer, C., Cuturi, M.: Missing data imputation using optimal transport. In: International Conference on Machine Learning, pp. 7130–7140. PMLR (2020)
21. Pereira, R.C., Santos, M.S., Rodrigues, P.P., Abreu, P.H.: Reviewing autoencoders for missing data imputation: technical trends, applications and outcomes. J. Artif. Intell. Res. **69**, 1255–1285 (2020)
22. Ranjbar, M., Moradi, P., Azami, M., Jalili, M.: An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. Eng. Appl. Artif. Intell. **46**, 58–66 (2015)
23. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
24. Song, Q., Shepperd, M.: Missing data imputation techniques. Int. J. Bus. Intell. Data Min. **2**(3), 261–291 (2007)
25. Stekhoven, D.J., Bühlmann, P.: Missforest-non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
26. Tashiro, Y., Song, J., Song, Y., Ermon, S.: CSDI: conditional score-based diffusion models for probabilistic time series imputation. In: Advances in Neural Information Processing Systems, vol. 34, pp. 24804–24816 (2021)
27. Van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. J. Stat. Softw. **45**, 1–67 (2011)
28. Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: International Conference on Machine Learning, pp. 5689–5698. PMLR (2018)
29. Zheng, S., Charoenphakdee, N.: Diffusion models for missing value imputation in tabular data (2023)