Online Imputation Techniques and Quality Assessment for Missing Values in Data Streams: A systematic Review

Software Project

Otto-von-Guericke Universität Magdeburg

Verfasser

Ahmed Hesham Kamal Khalifa (Matrikelnr: 231793)

Khaled Ahmed Ali Mobarak (Matrikelnr: 231935)

Noor Alsayed (Matrikelnr: 225867)

Hussein Al awassi (Matrikelnr: 235524)

Saden Bakrahji (Matrikelnr: 231685)

22.01.2024

Gutachter

Prof. Dr. Myra Spiliopoulou

Faculty of Computer Science (FIN), AG KMD: Wissensmanagement und Wissensentdeckung

Betreuer

M. Sc. Christian Beyer

Faculty of Computer Science (FIN) Institut für Technische und Betriebliche Informationssysteme (ITI)

Abstract

Addressing missing values in data streams presents a substantial hurdle across diverse domains like healthcare and social media analysis. Traditional imputation techniques, tailored for static datasets, often fall short when dealing with streaming data due to its real-time nature, incremental data arrival, and the occurrence of missing values at various time points. Consequently, it is crucial to provide effective and efficient imputation techniques that are tailored to particular requirements of data streams. To address this critical issue, we conducted a systematic review following the PRISMA 2020 guidelines to tackle two research questions (Q1) how missing values can be imputed in a data stream with low loss of information? (Q2) How can the quality or confidence of imputations be assessed in an online fashion? Our systematic review encompasses research studies that were published between 2019 and 2023. We utilized Google Scholar and ArXiv as our primary databases for this purpose.

During the initial phase, we identified 245 research papers by employing specific keywords for the first question, such as "Online data stream multiple imputations" and "Structurally missing data imputation," and for the second question, keywords including "Quality assessment of imputed data." To enhance the studies' quality, we filtered out studies containing surveys and non-English studies. Following a careful evaluation of the available research materials, we have retained 38 research papers for inclusion in the systematic review for both questions, adhering to the guidelines outlined in the PRISMA 2020 statement. In this review, we will discuss the various types of missing data, imputation techniques, and introduce novel methods for handling missing values in data streams. We constrained our search to a specific timeframe and a limited set of databases.

1 Introduction

In the rapidly evolving world of continuous data streams, the challenges posed by the abundance of data highlights the critical need for effective data imputation techniques. Missing data is a prevalent issue observed across diverse domains, including medicine, social sciences, and biology, among many others (Carpenter, James R., et al. (2023)). Various factors contribute to the emergence of missing values. Survey non-response, where participants choose not to answer specific questions, is one such factor. Additionally, missing data can arise from sensor malfunctions, drifts, network faults, or human errors during the data entry process (Ren, Lijuan, et al. (2014)).

While numerous studies have attempted to address data imputation challenges, many have not adequately grappled with the three V's, which constitute the primary challenges in data stream mining. These include efficiently processing and analyzing the high volume of data within limited time constraints, adapting to the rapid velocity at which data is generated, and managing the volatility inherent in the ever-changing data patterns and distributions, rendering it highly unpredictable (Krempl, Georg, et al. (2023)). As a result, recent research on data imputation has witnessed a clear increase.

This study aims to shed light on the most effective imputation techniques and methods for handling missing values in data streams, ensuring minimal information loss. Additionally, it seeks to provide insights into the assessment of imputation quality and confidence in an online fashion.

2 Methods

For our systematic review, we have followed the updated PRISMA 2020 guidelines (Sohrabi, Catrin, et al.), where we have abided the checklist in terms of eligibility criteria, information sources, search strategy, selection process, data collection process, synthesis, results, and discussion.

2.1 Eligibility/Inclusion Criteria

we have specified few inclusion and exclusion criteria to establish a clear and systematic process for selecting the studies to be included in our review. We have only included free full-text articles, as well as paper that require institutional access that are available in English, and we have systematically excluded paper that were published outside the timeframe spanning from 2019 to 2023. Moreover, we have excluded surveys.

2.2 Information Sources and Search Strategy

For our information Sources, we chose to perform searches in the following databases:

- Google Scholar
- ArXiv

To enhance the precision of our search strategy, we implemented an optimal filtering solution. This filtration process was executed by the authors of the research, aligning with our inclusion criteria. We created a keyword list for each research question. For both research questions we specified a combination of different keywords (Table 1).

Table 1: Combination of Keywords

Research question 1	**Research question 2**
Incremental imputation for data stream mining	Quality assessment of imputed data
Online data stream multiple imputations	Performance evaluation of prediction in imputation
Structurally missing data imputation	Prediction validation technique for imputation

We have initially identified 245 papers from the two distinct sources (Google Scholar and ArXiv). Authors SB and NA found 82 papers from Google Scholar, while author AK contributed an additional 58 papers, all selected based on keywords relevant to the first research question. Researchers KM and HA independently extracted 105 papers from arXiv, again focusing on the specific keywords associated with our research query.

From the collected pool of research papers sourced through these two platforms and keyword-driven selection, a filtering process was performed. At first, we used Zotero, a reference management tool, that manages and organizes our initial search results. Duplicate records were identified and removed using the deduplication features in Zotero. The remaining unique records were then subjected to the screening process. We then started the screening process, where we view and inspect the title and abstract, and decide whether the paper can be included and move to the second stage of selection. After removal of duplication and screening we have included 72 articles.

2.3 Selection and Data Collection Process

For data extraction and analysis, we identified and reviewed 72 articles and research papers through a thorough full-text search. These papers were then distributed among our team of researchers. While two researchers (NA, SB) collaborated and worked together as a team, three researchers (AK, KM, HA) worked independently on their assigned articles. During the extraction process, we came across occasional uncertainties regarding the information and data to derive from specific papers. Questions arose regarding these papers were about suitable imputation methods and whether these methodologies were compared to the state-of-the-art imputation methods. To address these discrepancies and uncertainties, we adopted a collaborative approach. When we faced ambiguities, we engaged in discussions as a team, deliberating on the papers and we all agreed on the extracted information. This collaborative work and effort ensured a more consensus-driven data extraction process, overcoming potential hurdles and enhancing the reliability of our findings. We included 24 research papers after the full-text search. We created a PRISMA 2020 flow diagram that shows the searches of databases and the included papers (Figure 1).

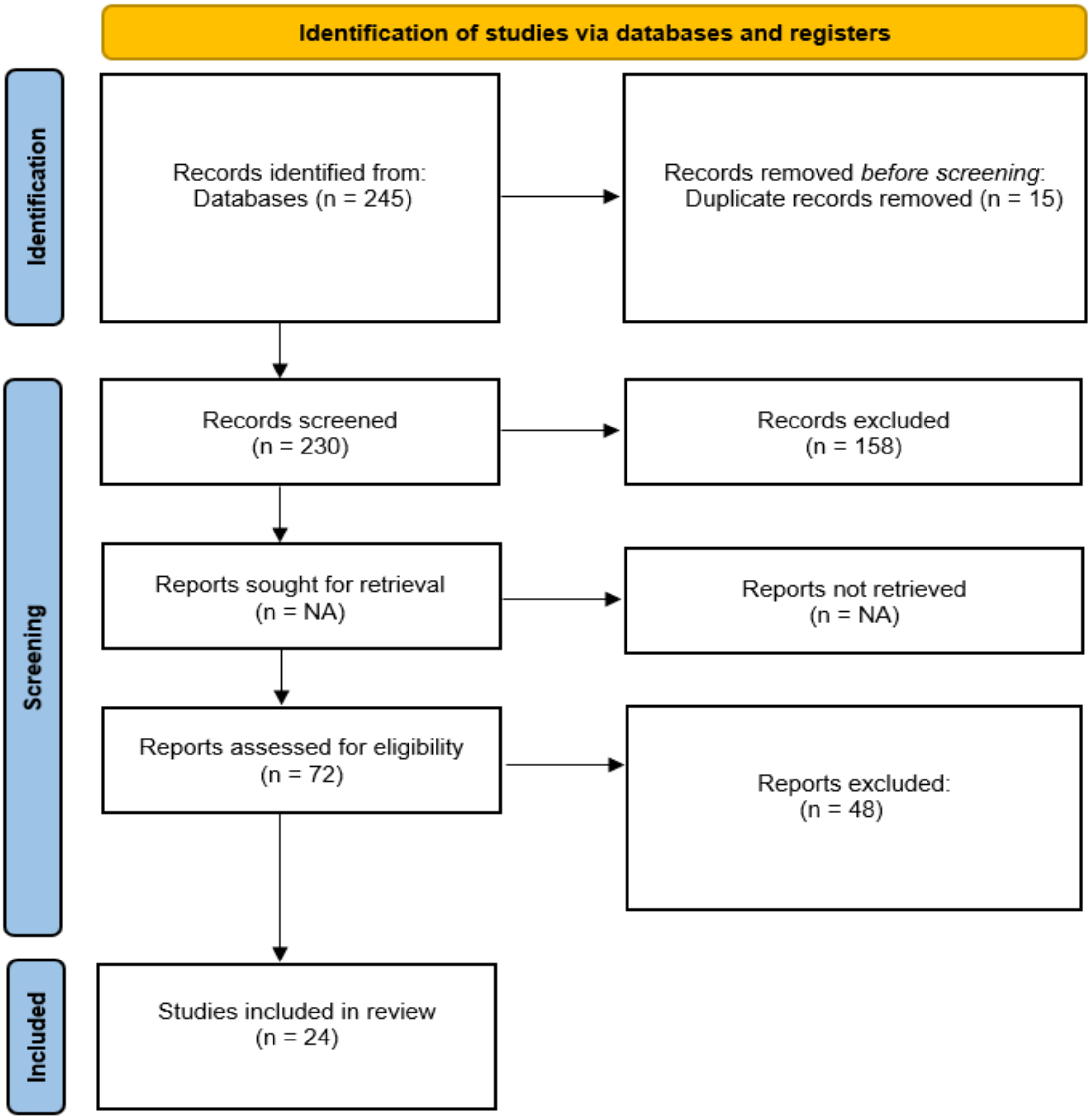

Figure 1: PRISMA flow diagram 1

To answer the second research question, we looked for relevant studies using specific keywords. We searched in both Google Scholar and ArXiv, like how we did for the first question. Initially, we found 180 research papers. We then deduplicated and screened these papers based on certain criteria, mirroring the methodology outlined for the first question. After this process, we ended up with 37 research papers that met our standards. To ensure precision of selected data, we refined the filtering even more, by full-text search resulting in a final set of 14 research papers that we decided to thoroughly examine (Figure 2).

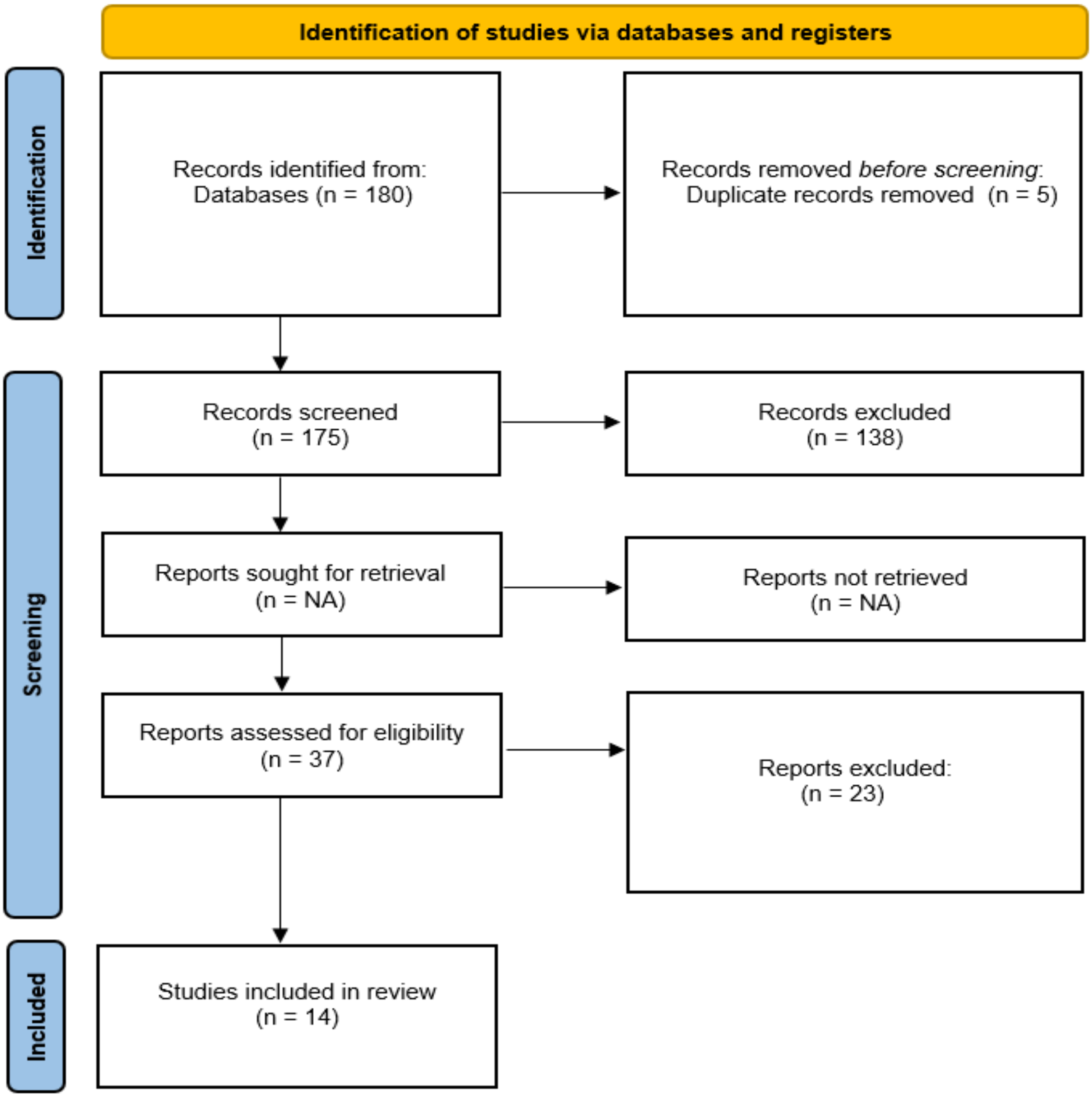

Figure 2:PRISMA flow diagram 2

2.4 Synthesis for Research Questions

Following the PRISMA guidelines, we conducted a thorough topic modeling analysis to synthesize the wide range of papers on multiple imputation techniques. We initially selected ten as the number of topics, which is a widely used initial choice in topic modeling, although its appropriateness can vary depending on the dataset or specific studies. To refine our model, we iteratively reduced the number of topics, striving to preserve essential themes while minimizing redundancy. Even after reducing the topics to 7, some overlapping persisted. Therefore, we continued refining our model, ultimately settling on a more focused representation including only 5 topics. The initial topics generated exhibited redundancy for each (supposedly) different topic. Even with 5 topics to be generated, we still observed substantial similarity and overlapping in the resulting words of each topic. This compelled us to investigate this issue further. Initially, our focus was on identifying the root problems before attempting to address an issue whose exact nature was not yet clear. During this exploration, we identified a couple of key challenges. Primarily, the challenge arose from a lack of diversity in our dataset. Given that our research addressed a specific question, the content within our datasets displayed a high degree of similarity. This similarity posed a difficulty in generating diverse and distinct topics from the corpus. While acknowledging the constraints in our ability to rectify these challenges outright, we undertook refinements to our algorithm to enhance its performance. A pivotal enhancement involved the implementation of lemmatization, a text normalization technique aimed at reducing words to their lexical or root form. This approach yielded several benefits, including dimensionality reduction, improved topic interpretation, consistent representation, and the mitigation of sparsity, where the robustness of the model is improved by lowering the number of unique terms.

Increasing the number of passes for the LDA Model emerged as a key factor in producing topics with enhanced distinctiveness and significance. This augmentation involves training the model on the entire corpus multiple times. However, it's important to exercise caution in this approach due to its associated trade-offs. While an increased number of passes offers advantages in terms of topic refinement, it concurrently amplifies computational time, requiring a balanced consideration of these factors.

We also enhanced the removal of irrelevant words from our text data; we initially employed the NLTK library, which includes a comprehensive set of commonly used stop words (Shu, X., & Cohen, R. (2010)). However, due to the nature of the studies included in our research, which often contain numerous numbers, equations, and formulas, we found it necessary to supplement NLTK's stop words with a manually curated list. This manual addition aimed to exclude specific figures that frequently appeared in the list of generated topics. This combined approach ensured a more thorough removal of unwanted elements, allowing our topic modelling process to focus on the content of the text.

Alpha and eta are crucial hyperparameters in LDA modelling. Alpha influences how documents distribute their preference over topics, while eta influences how topics distribute their preference over words. By default, both are set to a symmetric, 1 / num_topics prior. These parameters act as smoothing factors: higher alpha makes document preferences across topics smoother, and higher eta makes topic preferences across words smoother. They play a role in regulating the granularity of topic distribution in documents and word distribution in topics (Seth, N. (2021).).

We chose to set both alpha and eta to 'auto,' allowing the model to estimate these hyperparameters from the data rather than relying on manual specification. This strategy is advantageous because it makes the LDA model more adaptive, data-driven, and robust in capturing the underlying patterns of topics and words in your corpus. Gensim utilizes an empirical Bayes method to estimate these hyperparameters, leveraging the statistics derived from the observed data.

3 Results

3.1 Study Selection

As shown in figure 3, tackling the first research question, 24 papers from 2019 to 2023, were found through our search, showing a clear trend in the amount of research produced throughout this time. Four publications were found in 2019, Subsequently, in 2020, the number of publications remained consistent at four, reflecting a sustained interest. The year 2021 witnessed a notable increase to six publications, indicative of a growing body of literature. In 2022, there was a slight decrease to five publications, while in 2023, the trend stabilized with five publications.

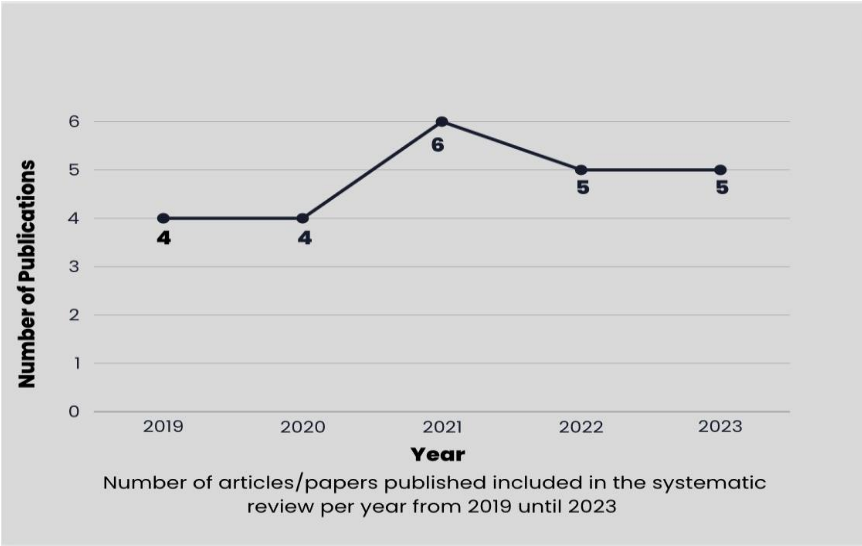

Figure 3: Papers published per year from year 2019 to 2023 RQ1.

The temporal representation of publications, essential for addressing our second research question, is shown in the following diagram. In our search within the specified timeframe of 2019 to 2023, we identified a total of 14 studies. There was a noticeable focus on the assessment of imputation techniques in 2019, with 5 publications contributing valuable insights to the field. The trend continued with 4 publications in 2020, followed by a slight decrease to 3 publications in 2021. Subsequently, in 2022, the number further decreased to 2 publications. Shockingly, the year 2023 exhibited a complete absence of publications meeting our criteria. Figure 4 illustrates the number of publications, highlighting a consistent decline in studies related to the assessment of imputation quality.

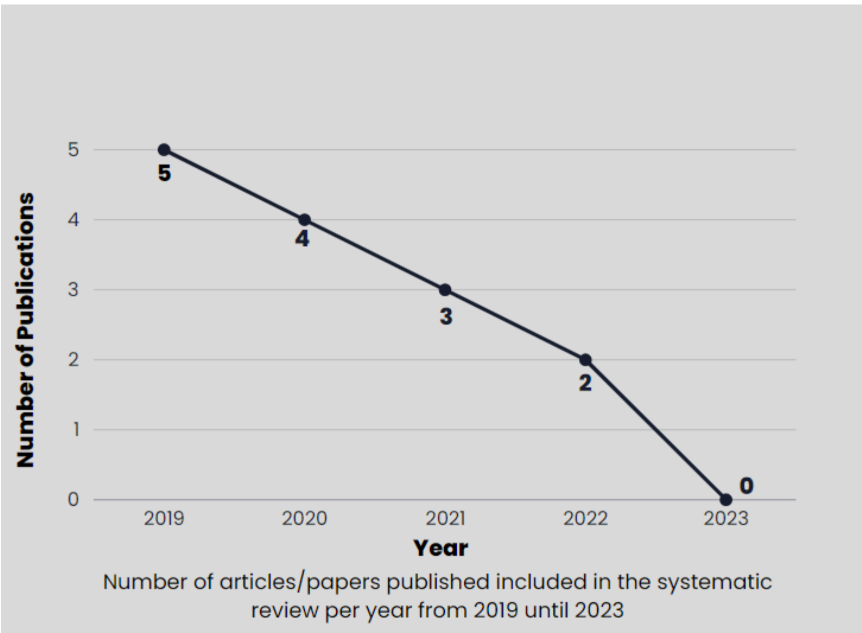

Figure 4: Papers published per year from year 2019 until 2023 RQ2.

3.2 Risk of Bias in Studies

Inapplicable in the concept of our study for both research questions, the concept of bias risk, as explained by Marshall et al. (2017), refers on the recruitment of participants in individual studies. The recruitment procedures in these studies hold no relevance to our inquiry since we do not aggregate the outcomes at the participant level.

3.3 Results of Synthesis

3.3.1 Results for RQ1:

All twenty-four studies included in this review were conducted between 2019 and 2023. Table 2 gives an overview of important characteristics of the selected studies. It shows the following, the novel or focused imputation model investigated, the methods used for comparative analysis, the corresponding ranking or evaluation outcomes, the specific evaluation metrics employed, the nature of the dataset (whether real-world or synthetic) and the type of missingness observed (Missing At Random - MAR, Missing Completely At Random - MCAR, or Missing Not At Random - MNAR).

Table 2: Summary of eligible publications RQ1

Research paper	Novel/focused model	Compared models	Ranking of models	Computational time (fastest to slowest)	Evaluation method	Datasets	Type of missingness
Chen, Katrina, et al. (2022)	GEDI	Mean imputation, kNN, MICE, SVD, GLFM, HIVAE, GRAPE	Error (GEDI, GRAPE, HIVAE, kNN, GLFM, MICE, SVD, MEAN) AUPRC (GEDI, GRAPE, MICE, HIVAE, kNN, GLFM, SVD, MEAN)	MEAN, SVD, GEDI, MICE, GRAPE, HIVAE, kNN=GLFM	RMSE AUPRC (for label prediction) Imputation time	Tabular classification datasets mixture of categorical and numerical features 9 datasets	MCAR
Lalande, Floria, and Kenji Doya (2023)	kNN X KDE	kNN Imputer, MISSFOREST, MICE, GAIN, SoftImpute, Mean, Median	kNN X KDE, MISSFOREST, kNN Imputer, MICE, GAIN, SoftImpute, Column Mean/Median	Small datasets (Mean, Median, MICE, kNN Imputer, kNN X KDE, SoftImpute, MissForest, GAIN) Large datasets 5000+ (Mean, Median, MICE, GAIN, SoftImpute, kNN Imputer, kNN X KDE, MissForest)	NRMSE, Time	3 synthetic datasets	Full MCAR, MCAR, MAR, MNAR
XUE, YE, et al. (2019)	MixMI/ MixMI-LL	GP, MTGP, M-RNN,	MixMI, MixMI-LL, 3D-MICE,	M-RNN,GP, MTGP, GMM, MixMI-LL,	MASE, Time	2 real datasets and 2	not mentioned

		GMM, MICE, 3D-MICE	MICE, GMM, GP, M-RNN, MTGP	MICE, MixMI, 3D-MICE		synthetic datasets	
Razavi-Far, Roozbeh, et al. (2022)	PSMI (Pooling)	ELMI, KNN, PCAI	Pooled PCAI, KNN, ELMI, Un-pooled PCAI, KNN, ELMI	not mentioned	NRMSE, ACC	Real dataset	MCAR
Khan, Shahidul Islam, and Abu sayed Md Latiful Hoque (2020)	SICE	Binary: MICE (PMM), FURIA, SVM Ordinal: MICE and SICE both using (PMM, POLYREG, CART, LDA) Numeric: SICE (BLR), MICE (PMM), MICE(BLR), Amelia, kNN	Binary: SICE, MICE, FURIA, SVM Ordinal: MICE and SICE both have similar performance Numeric: SICE (BLR), MICE (PMM), MICE(BLR), kNN, Amelia,	Ordinal: MICE using LDA is the fastest, SICE is always a bit slower. Numeric: Amelia, MICE (BLR), SICE (BLR), kNN, MICE (PMM)	Accuracy, Sensitivity, Precision, Specificity, F-measure, RMSE, Time	-) 4 data sets	MAR
Okafor, Nwamaka U., and Declan T. Delaney (2021)	VAE	NNRW, MICE, MISSFOREST, kNN	VAE, NNRW, MICE, kNN, MISSFOREST	not mentioned	RMSE	-) 2 real datasets	not mentioned
Karmitsa, Napsu, et al. (2020)	IVIACLR	MICE, Regression, Mean	U500: IVIACLR, MICE, Regression, Mean D500: Regression and Mean show advantage over MICE and IVIACLR U10000: IVIACLR, MICE Real datasets: IVIACLR, MICE, Regression, Mean	not mentioned	RMSE, MAE, UCE, CCD	-) 3 artificial -) 5 real datasets	MCAR, MAR, MNAR
Riggi, S., D. Riggi, and F. Riggi. (2020)	ML	Multiple imputation, Mean, ML-MN, ML-MSN	ML-MN, ML-MSN, MI, Mean	not mentioned	efficiency	Real dataset	MAR, MCAR

Dai, Zongyu, et al. (2023)	NNGP	MICE, GAIN, SoftImpute, Sinkhorn, Linear RR, MIWAE, Column Mean	NNGP, SoftImpute, MICE, MIWAE, Linear RR, Sinkhorn, Column Mean, GAIN	NNGP, SoftImpute, Sinkhorn, MICE, GAIN, MIWAE, Linear RR	MSE, Time	Synthetic and real datasets	MAR, MCAR
Spinelli, Indro, et al. (2020)	GINN/GNN	MICE, MIDA, RF, MissForest, Mean, Median, kNN	RMSE&MAE: GINN, RF, MissForest, kNN, MIDA Accuracy: GINN, MICE, kNN, Median, MissForest, RF, MIDA	not mentioned	MAE, RMSE, Accuracy	20 datasets	MCAR
Zhang, Xiaochaun, et al. (2023)	AmGCL	NeighAggre, VAE, GNN, GraphRNA, ARWMF, FP, SAT, SVGA	Recall&NDCG: AmGCL, SVGA, SAT, FP, GraphRNA, ARWMF, GNN,NeighAggre, VAE	AmCGL, SAT=SVGA	Recall, NDCG, Time	7 datasets	not mentioned
Wang, Ding, et al. (2023)	PoGEVON	Mean, Matrix Factorization (MF), MICE, BRITS, rGAIN, SAITS, TimesNet, GRIN, NET	PoGEVON, TimesNet, BRITS, GRIN, NET, rGAIN, MICE, SAITS, MF, Mean	not mentioned	MAE, MRE, MSE	5 Real-world Datasets	not mentioned
Kim, SeungHyun, et al. (2023)	supnotMIWAE	Mean, SAITS, Forward, GP-VAE	supnotMIWAE, Forward, GP-VAE, Mean, SAITS	not mentioned	MAE, MRE	3 real datasets	MNAR
Petrazzini, Ben Omega, et al. (2021)	—	KNN, MissForest, Amelia, MICE, MI, Mean	MissForest, KNN, MICE=Amelia, mi, Mean	Amelia, MICE, KNN, mi, MissForest	MAE, RMSE	31245 variants in the dataset	MAR, MCAR, MNAR
Liu, Shao-Hsien, et al. (2019)	IPW	MICE, Complete case analysis (CCA)	MICE, IPW=Complete case	not mentioned	MSE, Relative Bias	Simulated datasets	MAR, MCAR, MNAR
Zhao, Yuxuan, et al. (2022)	Online EM	minibatch EM, GROUSE,	Offline EM, Minibatch EM,	GROUSE, KFMC, Minibatch EM,	MAE, RMSE, Runtime	Offline and Online real	MCAR

		KFMC, Offline EM	Online EM, KFMC, GROUSE	Online EM, Offline EM		data experiments	
Hamzah, Fatimah Bibi, et al. (2021)	RRRI	KNN, CART	RRRI, KNN, CART When combined with MLR (RRRI-MLR, CART-MLR, KNN-MLR)	not mentioned	CE, RMSE, MAPE	Streamflow datasets	not mentioned
Kunicki, Robert, and Maciej Grzenda. (2021)	UTRIIDS as an imputation method assigned to ML methods	Naïve Bayes, ARF, KNN, Hoeffding Tree, HAT	Naïve Bayes, KNN, ARF, HAT, Hoeffding Tree	not mentioned	Average k coefficient	6 datasets	not mentioned
Madley-Dowd, Paul, et al. (2019)	MI	CCA	MI, CCA	not mentioned	FMI, empirical SE	1000 Simulated dataset	MAR, MCAR
Zhang, Yifan, and Peter J. Thorburn. (2022)	Dual-SSIM	MICE, Mean, LOCF, Linear, EM, KNN, SSIM, BRITS, M-RNN	Dual-SSIM, BRITS, SSIM, M-RNN, LOCF, MICE, Mean, Linear, EM, KNN	not mentioned	RMSE, MAE	2 real datasets	MAR
Zhu, Xiaofeng, et al. (2019)	DIM	Random Imputation (RI), Incremental Imputation Algorithm (IIA), Maximal Economics (ME), Least Economics (LE), Mutual information	DIM, Mutual Information, LE, ME, IIA, RI	not mentioned	PA, CA	6 real datasets	not mentioned
Lim, David K., et al. (2021)	NIMIWAE	HIVAE, IMIWAE, VAEAC, MIWAE, MICE, Mean, MissForest	NIMIWAE, IMIWAE, VAEAC, MICE, MissForest, HIVAE, MIWAE, Mean	not mentioned	Average L1 distance, Percent Bias	1 Real dataset	MCAR, MAR, MNAR
Lin, Yiming, and Sharad Mehrotra. (2022)	QUIP	KNN, XGBoost, Mean, LOCATER, ImputeDB	QUIP Lazy (assigned to KNN, XGboost) QUIP Adaptive (Mean, Locater) outperforms ImputeDB	QUIP Lazy, QUIP Adaptive, ImputeDB	Running time	2 real datasets and 1 synthetic	not mentioned

Dong, Wenlu, et al. (2021)	(AS, WAS) combined with other methods	Mean, KNN, BPCAI	AS-BPCAI, AS-KNN, WAS-Mean, AS-Mean, WAS-KNN	not mentioned	RMSE	3 synthetic datasets and 2 real-world datasets	not mentioned

One interesting finding in the previous corpus of the 24 papers we included is that there was a strong emphasis on new imputation models. Of the 24 papers that were examined 23 papers-the vast majority-introduced novel methods of imputation. Whereas, the study conducted by Petrazzini, Ben Omega, et al. (2021) stands out as it did not introduce a novel imputation technique; instead, it focused on a comparison of existing imputation methods. We identified a pool of 68 different imputation techniques that could be useful comparison models in order to thoroughly assess and compare these unique models. 52 imputation techniques were used only once across the several papers, highlighting the diversity in the approaches considered. However, a subset of 16 techniques emerged as frequent contenders, indicating their prevalence in the literature. The information detailing the frequency of usage among the encountered imputation techniques is summarized below and is visually represented in Figure 3. The techniques Mice and Mean were particularly prevalent, each mentioned in 54.2% of the papers, followed closely by KNN with 50%. Missforest, while less frequent, still appeared in 20.8% of the papers, demonstrating its relevance. A set of 12 additional techniques, including MIWAE, HIVAE, M-RNN, BRITS, EM, CART, CCA, Amelia, SAITS, Median, Softimpute, and GAIN, each occurred in 8.3% of the papers or two times each among all the 24 papers.

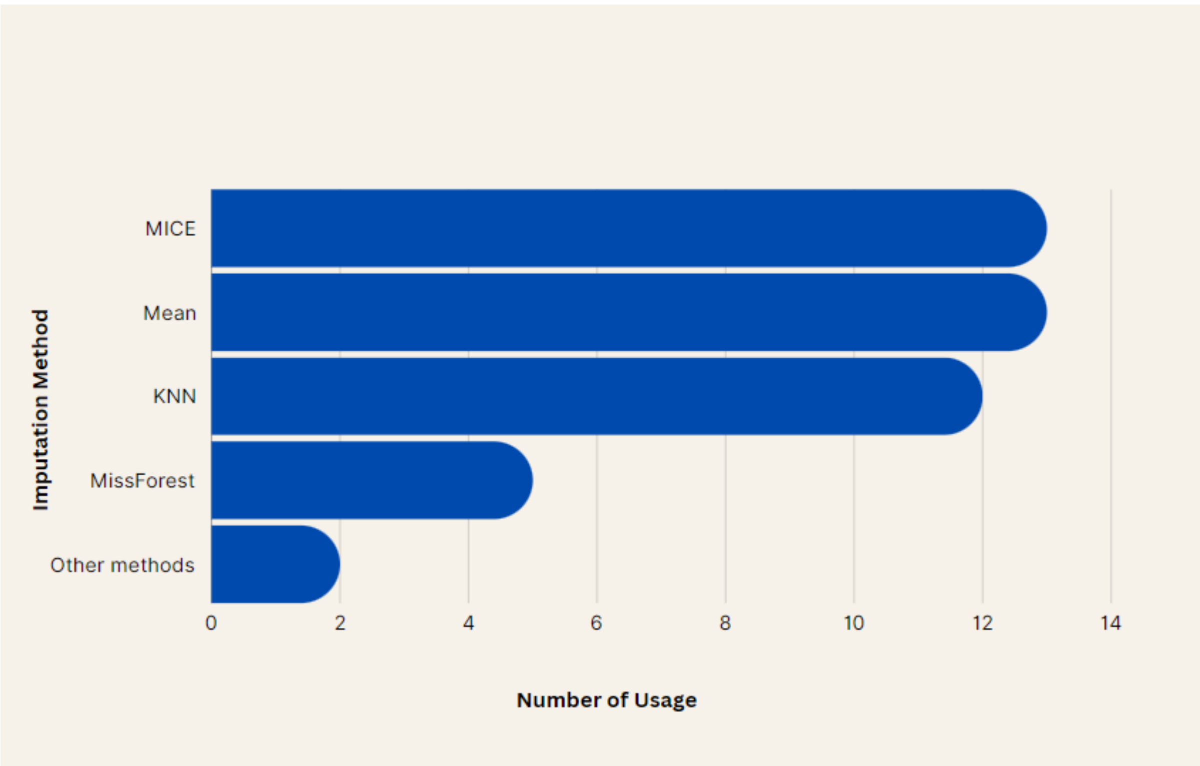

Figure 5: Number of Usages for Different Imputation Methods

Table 3 represents an in-depth overview of numerous novel imputation methods along with explanations of their distinct objectives and outcomes. It serves as a valuable resource for understanding the wide range of imputation techniques.

Table 3: Novel Imputation Techniques

Novel model	Concept and objectives	Outcome
GEDI (Graph and Transformer-based Data Imputation)	Handling missing data in tabular datasets. The model aims to "preserve both row-wise similarities among observations and column-wise contextual relationships among features in the feature matrix" and tailoring imputation to downstream tasks.	Directly utilize downstream information, that enhances the efficiency of the label prediction. Outperforms all compared models.
kNN X KDE (k-nearest neighbors X Kernel Density Estimation)	A hybrid imputation method that combines the k-nearest neighbor and Kernel Density Estimation to improve the accuracy of the imputation.	KNN X KDE preserves the actual data structure, it has achieved the best average imputation NRMSE in all data scenarios, however it becomes computationally expensive when the dataset is very large.
MixMI (mixture-based multiple imputation)	Imputation for both cross-sectional information and temporal correlations. Using Linear regression for cross-sectional information and Gaussian processes for temporal correlations. Training of model using Expectation maximization.	MixMI works on both cross-sectional and temporal correlations in time series. IT outperforms other state-of-the-art methods in accuracy; however, it is a bit time consuming.
SICE (Single Center Imputation from Multiple Chained Equations) Categorical and numerical	It is an extension of the MICE. Where MICE is used and repeated m times added to an array, then the missing value is substituted by the mean of its matching imputed value from the array.	SICE performs better than MICE in numeric data, however it doesn't show better performance in ordinal data. It achieved 20% improve in F-measure and 11% error reduction.
IVIACLR (Imputation via Clusterwise Linear Regression)	Clusterwise Linear Regression used to predict suitable imputation.	IVIACLR performs efficiently when data have very clear cluster structure specifically in MAR and MCAR data.
NNGP (Neural Network Gaussian Process)	Proper statistical inference and to perform well in high dimensional settings	a well-developed NNGP imputation model for high dimensional incomplete data, that is also robust to high missing rates
GINN (Graph Imputation Neural Network)	A GNN encoder creates intermediate representations by combining projection layers and local neighbour information. The decoding GNN reconstructs the imputed dataset. To enhance training speed and performance, various losses are employed, including Wasserstein adversarial loss with gradient penalty. In short, a novel graph convolutional autoencoder reconstructs the entire dataset.	The algorithm exhibits robustness to external classifier choices and outperforms competitors in experiments with high artificial noise levels. While not consistently achieving the top accuracy, it demonstrates superior resilience across classifiers.
AmGCL (Attribute missing Graph Constrastive Learning)	amGCL is a graph neural network model specifically created to tackle the issue of missing attribute data in graphs. The model	AmGCL surpasses the other methods in terms of training time. Experimental findings on various real-world datasets

	utilizes self-supervised graph augmentation contrastive learning to enhance its performance	highlight AmGCL's superior performance in feature imputation and node classification compared to state-of-the-art methods.
PoGEVON (Position-aware Graph Enhanced Variational Autoencoders)	PoGeVon utilizes a variational autoencoder (VAE) to predict missing values across both node time series features and graph structures.	PoGeVon consistently outperforms strong baseline methods in imputing missing values for node time series across various real-world datasets.
IPW (Inverse Probability Weighting)	IPW simplifies the application of MSMs by employing a direct approach, using logistic regression to calculate inverse probability weights for observed treatment or censoring. The method focuses on estimating parameters in the context of incomplete data by assigning weights based on the probability of having complete data for each participant.	MI seems advantageous over IPW in MSMs applications, with the former providing consistent empirical power across scenarios. While IPW concentrates on predicting missing data mechanisms and may outperform MI in certain situations, the MI approach holds an advantage in MSMs applications, particularly under realistically constructed scenarios.
CCA (Complete Case Analysis)	statistical analysis solely involves participants with complete data on the variables of interest, excluding those with any missing data.	CCA exhibited a pattern of reduced power as the proportion of missing data increased.
RRRI (Robust Random Regression Imputation)	RRRI represents a less rigid version of least squares regression, functioning with more relaxed assumptions. It provides notably improved estimations of regression coefficients in scenarios where the data are uncertain.	The RRRI method had the highest CE and the lowest RMSE and MAPE values.
SSIM (Sequence-to-Sequence Imputation Model)	SSIM, the initial data imputation model employing sequence-to-sequence architecture and attention mechanism, utilizes LSTM to capture temporal information between gaps. The global attention mechanism allows SSIM to concentrate on specific input segments when estimating various missing values.	The next best imputation models after Dual-SSIM are neural network-based methods like SSIM, BRITS and M-RNN which outperform both the statistical and model-based solutions significantly.
QUIP (query-time missing value Imputation)	QUIP is a query-time approach for imputing missing values, leveraging query semantics to minimize cleaning overhead.	Actual experiments demonstrate that QUIP surpasses the state-of-the-art technique ImputeDB by a factor of 2 to 10 times, achieving a significant improvement over conventional offline approaches in terms of order of magnitudes.
NIMIWAE (Non-Ignorable Missing Data using Importance-Weighted Autoencoders)	considers missing observations as latent variables within the VAE framework, employing Importance-Weighted	NIMIWAE adeptly handles MNAR missingness in VAE/IWAE methods for complex Physionet EHR data. Through

	Autoencoders (IWAEs). learns a valuable lower-dimensional data representation for tasks like patient subgroup identification and data visualization.	simulations, it excels in imputing features under MNAR and performs well under MCAR/MAR.
DIM (Date-driven Incremental Imputation Model)	DIM utilizes all available information in the dataset for economical, effective, and iterative missing value imputation. The aim of DIM is to orderly impute missing values, minimizing imputation costs and maximizing accuracy. Specifically, DIM identifies unnecessary imputation for absent and predictable samples to reduce cost and noise.	DIM consistently outperforms MI, ME, LE, IIA, and RI in both prediction and classification accuracy across various missing rates.
supnotMIWAE (supnot-Missing data Importance-Weighted AutoEncoder)	A probabilistic framework for multivariate time series data with missing values.	The method outperformed the dedicated imputation baseline (SAITS) with the lowest MAE and MRE, surpassing the GP-VAE model as well. Despite these achievements, the forward imputation structure has limitations in handling diverse time series patterns, especially those with sudden spikes or periodicity.

3.3.1.1 Overview of state-of-the-art imputation methods

In this section, we listed and described the leading state-of-the-art imputation methods, which served as benchmarks for the evaluation of novel techniques.

I. Multiple Imputation by Chained Equation (MICE)

The MICE method is an imputation technique commonly used to address missing data in datasets. MICE imputes missing values by iteratively modelling each incomplete variable conditional on the others. First, all missing values are filled by random sampling data like mean value imputation, this acts as a place holder. Then the place holder for one of the variables (X_i) is set back to missing. For each variable with missing value, use linear regression with other variables. Then observed values are used in the model, with the variable of interest serves as the dependent variable. Finally, the missing values are replaced with the predictions derived from the regression model. The process is then repeated several times for a refined result (Okafor, Nwamaka U., and Declan T. Delaney (2021)).

II. Mean

Mean imputation is a simple method for handling missing data. The basic idea is to replace the missing values with the average value of the observed data. At the beginning, we identify the variables that have missing values, then for each of the missing values, the mean of the observed values of this variable is calculated $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. After that, the missing value in each variable is substituted with the calculated mean for that variable. Finally, process is repeated for all variables with missing values (Zhang, Yifan, and Peter J. Thorburn. (2022)).

III. K-Nearest Neighbor (KNN)

KNN is often used as a method to estimate missing values based on the value of their nearest neighbors in a dataset. For each missing value in a dataset, identify the k non-missing values that are closest to the missing data point. Subsequently, the distance between the missing point (x) and each non-missing neighbor (y) is calculated using the Euclidean distance dist(x_i,y_i) = $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$, then the average of the values from the k closest neighbors is taken and used to fill the missing point. Process is then repeated for each variable with missing values (Okafor, Nwamaka U., and Declan T. Delaney (2021)).

IV. MissForest

Missforest uses an iterative method based on Random Forest algorithm. The RF algorithm is trained in the observed values of the dataset, which has an in-built mechanism for handling missing data. It weighs the frequency of observed values on a variable with the RF proximities after being trained on an initially mean-imputed dataset. Then an iterative imputation process begins, the dataset is separated into observed and missing parts for each variable. An initial guess is made for the missing value using the mean imputation. Then the variables are sorted based on the number of missing values, starting with the variable with the lowest number of missing values. For each variable with missing values, an RF model is fitted with observed values as the response and other observed variables as predictors. The missing values are then predicted by applying the trained RF model to the corresponding set of missing values. Finally, and early stopping criterion is set to avoid overfitting and the imputation iteration is repeated until it reaches this criterion (Okafor, Nwamaka U., and Declan T. Delaney (2021)).

3.3.1.2 Type of Missingness

After the missingness patterns from all 24 research publications were analyzed, many patterns in the data characterization were found. These patterns showed different percentages of cases of Missing At Random (MAR), Missing Completely At Random (MCAR), and Missing Not At Random (MNAR). An important finding was that 12 papers predominantly reported the presence of Missing Completely At Random (MCAR) data, where every measurement in the dataset has the same probability of being missing, and the causes of the missing data are unrelated to the data. This assumption implies that the missing values occur randomly and independently of any observed or unobserved data. On the other hand, Missing At Random (MAR) patterns were found in 10 publications, indicating that only groups of measurements in the datasets have the same probability of being missing, and the observed data define this probability. MAR is considered a more general and realistic assumption than MCAR, allowing the missingness to be modeled using the observed data. Finally, 6 papers recognized the presence of Missing Not At Random (MNAR) data, which indicates that the probability of data being missing is related to unobserved factors or variables (Zhang, Yifan, and Peter J. Thorburn. (2022)). The distribution of these missingness patterns highlights how crucial it is to understand the nature of missing data in the context of imputation techniques. This helps in the decision of choosing the right method based on the dataset´s observed characteristics.

In parallel with our analysis of missing data patterns, we have also examined the evaluation metrics applied across the included studies. The frequency of metric usage provides important information about the different aspects used for evaluating and assessing the imputation methods. RMSE (Root Mean Squared Error) was used 10 times across the 24 studies, accordingly, it is the most frequently utilized metric. Computational time attracted a lot of attention, featuring in evaluations across 8 papers. Followed by MAE (Mean Absolute Error) with 7 instances. Additionally, various metrics, such as NRMSE, MSE, MRE, ACC, and others, were utilized with varying frequencies. Table 4 presents detailed information on each metric, including equations, definitions, and number of times the metric is used.

Table 4: Evaluation Metrics 1

Evaluation Metric	Equation	Frequency of Use
RMSE (Root Mean Squared Error)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}({y_i}^{obs} - {y_i}^{imp})^2}$	10
Computational Time	—	8
MAE (Mean Absolute Error)	$\frac{1}{n}\sum_{i=1}^{n}(\lvert {y_i}^{obs} - {y_i}^{imp}\rvert)$	7
MSE (Mean Squared Error)	$\frac{1}{n}\sum_{i=1}^{n}(\lvert {y_i}^{obs} - {y_i}^{imp}\rvert)^2$	3
ACC (Accuracy)	$\frac{True\ Positives + True\ Negatives}{N}$	3
NRMSE (Normalized Root Mean Squared Error)	$\sqrt{\frac{1}{N_{miss}}\sum_{i=1}^{N}\sum_{j=1}^{D}(X_{ij} - x_{ij})^2(1 - m_{ij})}$ Where $N_{miss} = \sum_{i=1}^{N}\sum_{j=1}^{D}(1 - m_{ij})$	2
MRE (Mean Relative Error)	$\frac{1}{n}\sum_{i=1}^{n}\lvert\frac{{y_i}^{obs} - {y_i}^{imp}}{{y_i}^{obs}}\rvert$	2
Sensitivity/Recall	$\frac{True\ Positives}{True\ Positives + False\ Negatives}$	2
AUPRC (Area Under Precision-Recall Curve)	—	1
UCE (Unsupervised Classification Error)	% of misclassified error	1
CCD (Cluster Center Displacement)	$\frac{1}{k}\sum_{i=1}^{k}(\lvert\lvert {c_i}^{obs} - {c_i}^{imp}\rvert\rvert)$	1
NDCG (Normalized Discount Cumulative Gain)	$\frac{DCG_K}{IDCG_K}$	1

FMI (Fraction of Missing Information)	$\frac{B}{(W+B)}$ B= $((\frac{1}{m-1})\sum_{k=1}^{m}(\widehat{\beta_k}-\hat{\beta})^2)$ W = (1/m) $\sum_{k=1}^{m}\widehat{V_k}$	1
Empirical SE (empirical Standard Error)	$\sqrt{(\frac{1}{r-1})\sum_{q}^{r}(\widehat{\beta_q}-\bar{\beta})}$	1
CE (efficiency Coefficient)	1-$\frac{\sum_{i=1}^{n}(Y_i-y_i)^2}{\sum_{i=1}^{n}(Y_i-y_i)^2}$	1
MAPE (Mean Absolute Percentage Error)	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert {y_i}^{obs}-{y_i}^{imp}\rvert}{{y_i}^{obs}}$	1
PA (Prediction Accuracy)	$\frac{1}{t}\sum_{i=1}^{t}l(IV_i,RV_i)$	1
CA (Classification Accuracy)	$\frac{1}{n}\sum_{i=1}^{n}l(IC_i,RC_i)$	1
MASE (Mean Absolute Scaled Error)	$\frac{1}{\sum_p I_{p,v}}\sum_{P}\frac{\sum_{i\in mask_{p,v}}\lvert x_{p,v,i}-X_{p,v,i}^{obs}\rvert}{\frac{J_{p,v}}{J_{p,v}-1}\sum_{j=2}^{J_{p,v}}\lvert Y_{p,v,j}-Y_{p,v,j-1}\rvert}$	1
Precision	$\frac{True\ Positives}{True\ Positives+False\ Positives}$	1
Specificity	$\frac{True\ Negatives}{False\ Positives+True\ Negatives}$	1
F1-measure	$2*\frac{Precision*Recall}{Precision+Recall}$	1
PB (Percent Bias)	$\frac{1}{p}\sum_{j=1}^{p}\frac{\lvert\beta_j-{\beta^{\wedge}}_j\rvert}{\lvert\beta_j\rvert}$	1
RB (Relative Bias)	$(\frac{\beta-\beta_{truth}}{\beta_{truth}})$*100%	1
Average L1 distance	$\frac{\lvert X^m-x^x\rvert}{N_{miss}}$	1
Average K coefficient	—	1
Efficiency	—	1

3.3.2 Results for RQ2:

In addressing the second RQ, our focus centers on the evaluation metrics deployed for assessing the quality or confidence of imputations in an online fashion. We have created a table that has the same format used for the first RQ, where it summarizes the information gathered from the 14 studies included in the systematic review for RQ2. The emphasis here is placed on elucidating the diverse evaluation methods rather than the novel imputation techniques themselves. The data is presented in Table 5, However we have removed the computational time, since it adds no benefit to answer the research question.

Table 5: Summary of studies RQ2

Research paper	Novel/focused model	Compared models	Ranking of models	Evaluation method	Datasets	Type of missingness
Shadbahr, Tolou, et al. (2022)	—	MICE, MissForest, MIWAE, GAIN, Mean	MissForest, GAIN, MICE, MIWAE, Mean	RMSE, MAE, R^2, KL divergence, KS statistic, 2-wasserstein distance, Sliced Wasserstein distance	Simulated and real datasets	MCAR
Xu, Xiaoting, et al. (2022)	—	Simple imputation, KNN, SVD, iterative imputation	Iterative, KNN, Simple, SVD	SMAPE, F1-score, ACC	Real-world dataset	not mentioned
Pourshahrokhi, Narges, et al. (2021)	F-HMC	MICE, KNN, PPCA, MissForest	F-HMC, PPCA, MICE, KNN, MissForest	NRMSE, Precision, ACC, Recall, F-1 score	Real-world dataset	not mentioned
Mera-Gaona, Maritza, et al. (2021)	MICE	Basic Imputation method	MICE, Basic Imputation method	MAE, RMSE, ACC	Real-world dataset	not mentioned
Cheng, Ching-Hsue, Chia-Pang Chan, and Yu-Jheng Sheu. (2019)	PKNNI	AI, ZI, CMI, KNNI, MI	PKNNI, MI, KNNI, ZI, AI,CMI	ACC	8 UCI-datasets	MAR, MCAR, MNAR
Zhao, Junhui, et al. (2020)	DLIP	Mean	DLIP, Mean	MAE, MRE, NMSE	PeMS	MAR, MNAR
Hamzah, Fatimah Bibi, et al. (2021)	RRRI	KNN, CART	RRRI, KNN, CART When combined with MLR (RRRI-MLR, CART-MLR, KNN-MLR)	CE, RMSE, MAPE	Streamflow datasets	MAR
Afrifa-Yamoah, Eben, et al. (2020)	ARIMA	Multiple linear regression, Structural time series models	Multiple linear regression, ARIMA, Structural time series models	MAE, RMSE, SMAPE	Time series dataset	MAR
Kim, Taeyoung, Woong Ko, and Jinho Kim. (2019)	—	LI, Mode, KNN, MICE	KNN, MICE, LI, Mode	RMSD, RMSE, MRE, MRD	Solar power generation dataset	MCAR
Teh, Hui Yie, Andreas W. Kempa-Liehr,	—	fuzzy C-means clustering, KNN, Singular value	—	Recall, FPR, FNR; Precision, ACC, F-Score,	—	—

and Kevin I-Kai Wang. (2020)		decomposition, PMF		MCC, RMSE, MSE, MAE, MRE		
Vazifehdan, Mahin, Mohammad Hossein Moattar, and Mehrdad Jalali. (2019)	Hybrid imputation between Bayesian network model and tensor factorization	Mean/Mode, Hot-deck, KNN, Weighted KNN, Tensor model, Bayesian network model	Hybrid imputation, Bayesian network model, Tensor model, Mean/Mode, W-KNN, KNN, Hot-deck	NRMSE, ACC, Sensitivity, Specificity	Real-world dataset	MAR, MCAR, MNAR
Garcia, Cristiano, Daniel Leite, and Igor Škrjanc. (2019)	eFGP	eGNN, eTS, xTS, FBeM	eFGB, eXTS, eGNN, FBeM, eTS	RMSE, NDE	Real-world dataset	MCAR, MAR
Shi, Shuo, et al. (2019)	—	Beagle4.1, IMPUTE2, MACH+Minimac3, and SHAPEIT2+IMPUTE2	IMPUTE2, SHAPEIT2+IMPUTE2, MACH+Minimac3, Beagle4.1	Sensitivity, FPR, R^2	Real-world dataset	not mentioned
Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. (2020)	SICE	Binary: MICE (PMM), FURIA, SVM Ordinal: MICE and SICE both using (PMM, POLYREG, CART, LDA) Numeric: SICE (BLR), MICE (PMM), MICE(BLR), Amelia, kNN	Binary: SICE, MICE, FURIA, SVM Ordinal: MICE and SICE both have similar performance Numeric: SICE (BLR), MICE (PMM), MICE(BLR), kNN, Amelia,	ACC, Sensitivity, Precision, Specificity, F-measure, RMSE, Time	-) 4 data sets	MAR

Despite having entirely different keywords, two studies, Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. (2020) and Afrifa-Yamoah, Eben, et al. (2020) were identified in both searches conducted for RQ1 and RQ2. This discovery indicates that these studies tended to offer valuable insights to our review.

Table 6 focuses on the evaluation metrics used in the studies included to assess the quality of the imputation method and how accurate the imputed data is as in Table 4. The next discussion in section 4.1 will examine these metrics´ distinctions, clarifying their variations and looking into possible ways they could be combined with other methods. Through the analysis we aim to reveal strategies for optimizing and maximizing accuracy in imputation methods.

Table 6: Evaluation Metrics 2

Evaluation Metric	**Equation**	**Frequency of Use**
RMSE (Root Mean Squared Error)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}({y_i}^{obs} - {y_i}^{imp})^2}$	8

ACC (Accuracy)	$\frac{True\ Positives + True\ Negatives}{N}$	7
MAE (Mean Absolute Error)	$\frac{1}{n}\sum_{i=1}^{n}(\lvert y_i{}^{obs} - y_i{}^{imp}\rvert)$	5
Sensitivity/Recall	$\frac{True\ Positives}{True\ Positives + False\ Negatives}$	5
Precision	$\frac{True\ Positives}{True\ Positives + False\ Positives}$	3
MRE (Mean Relative Error)	$\frac{1}{n}\sum_{i=1}^{n}\lvert\frac{y_i{}^{obs} - y_i{}^{imp}}{y_i{}^{obs}}\rvert$	3
NRMSE (Normalized Root Mean Squared Error)	$\sqrt{\frac{1}{N_{miss}}\sum_{i=1}^{N}\sum_{j=1}^{D}\left(X_{ij} - x_{ij}\right)^2(1 - m_{ij})}$ Where N_{miss} = $\sum_{i=1}^{N}\sum_{j=1}^{D}(1 - m_{ij})$	2
Specificity	$\frac{True\ Negatives}{False\ Positives + True\ Negatives}$	2
F1-measure	$2 * \frac{Precision * Recall}{Precision + Recall}$	2
R^2 (coefficient of determination)	—	2
SMAPE (Symmetric Mean Absolute Percentage Error)	$\frac{\sum_{t=1}^{T}\lvert(\widehat{y_t} - y_t)/y\rvert}{T}$*100	2
FPR (False Positive Rate)	$\frac{False\ Positives}{True\ Negatives + False\ Positives}$	2
FNR (False Negative Rate)	$\frac{False\ Negatives}{True\ Positives + False\ Negatives}$	1
MCC (Matthew's Correlation Coefficient)	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	1
NDE (Normalized Deviation Error)	$\frac{RMSE}{std(y^{[h]}\forall h)}$ Where std is the standard deviation	1
MRD (Mean Relative Deviation)	$\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert\widehat{P_i} - P_i\rvert}{P_{total}} * 100(\%)$	1
RMSD (Root Mean Square Deviation)	$\sqrt{\sum_{i=1}^{N}\frac{1}{N}(\widehat{P_i} - P_i)^2}$	1
MAPE (Mean Absolute Percentage Error)	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i{}^{obs} - y_i{}^{imp}\rvert}{y_i{}^{obs}}$	1
CE (efficiency Coefficient)	1-$\frac{\sum_{i=1}^{n}(Y_i - y_i)^2}{\sum_{i=1}^{n}(Y_i - y_i)^2}$	1
NMSE (Normalized Mean Squared Error)	$\frac{\sum_{i=1}^{n}\lvert x_i - \hat{x}_i\rvert^2}{\sum_{i=1}^{n}x_i^2}$	1
MSE (Mean Squared Error)	$\frac{1}{n}\sum_{i=1}^{n}(\lvert y_i{}^{obs} - y_i{}^{imp}\rvert)^2$	1

In our exploration of evaluation metrics across the 14 studies, a diverse array of 21 metrics was employed to measure the performance of imputation methods. Notably, RMSE took the lead, being mentioned in 8 instances, being the frequently adopted benchmark. Following closely, Accuracy was used 7 times, while MAE and Recall emerged with 5 mentions each. Moreover, MRE and Precision appeared 3 times each. Beyond these key metrics, an additional 15 evaluation methods made appearances, each was used once or twice across the 14 studies.

4 Discussion

4.1 Discussion on RQ1

In this systematic review, we have precisely examined plenty of new and innovative imputation techniques, comparing them to established state-of-the-art methods. Our main goal was to assess the quality of the imputation techniques and determine their effectiveness in imputing missing values with minimal information loss. As shown in the results section 3.3, our analysis involved a comparison with leading benchmark methods such as MICE, KNN, MissForest, and Mean imputation. The findings from our review indicates that the novel imputation techniques consistently outperform the benchmark methods they are compared to. Across a range of evaluation metrics and diverse datasets, these novel methods show a remarkable ability to handle missing values with precision and efficiency.

Nevertheless, the superior performance of the novel methods, it is essential to recognize that some studies use synthetic datasets for their assessments. For instance, 14 synthetic datasets were used across all the studies, this type of datasets are valuable for controlled experiments, and they are designed with a specific purpose and might lack some of the intricacies and complexities present in real-world data (Figure 6). These synthetic datasets enable researchers to evaluate imputation approaches in certain scenarios and in specific domains.

However, most of the datasets (45) that exists in the studies are real-world datasets, providing a more accurate representation of the difficulties and complexities associated with missing values. Real-world datasets represent the actual data observed in practical scenarios, introducing a level of complexity that synthetic datasets may lack. The limitations of synthetic datasets, with their controlled and often simplified structures, could influence the performance metrics of imputation techniques. There is a degree of doubt regarding the applicability of the reported results due to the mismatch between synthetic and real-world data. Therefore, while these novel methods show outstanding performance on synthetic datasets, their real-world efficacy may be subject to different challenges.

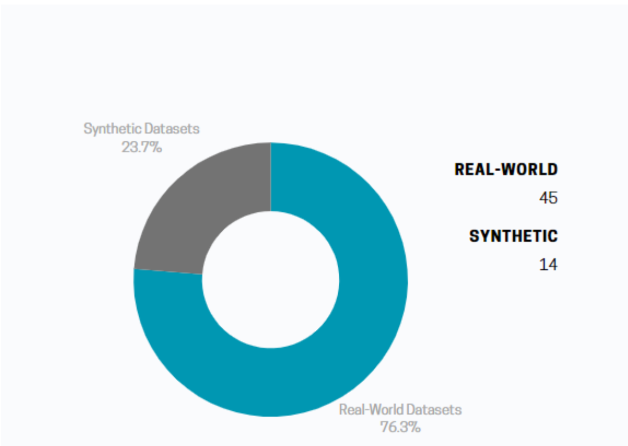

Figure 6: Difference between Real-world and Synthetic Datasets

Furthermore, the efficacy of imputing missing values with minimal information loss is linked to the nature of the data itself. The type of data, whether numerical, categorical, or specific to a particular domain (such as genetic data), significantly influences imputation outcomes. This variability highlights the fluctuations observed in the performance of methods like MICE, which may rank as the second-best method as in Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. (2020) and one of the least effective in Wang, Ding, et al. (2023).

4.2 Discussion on RQ2

Evaluating the quality and how good does an imputation method perform is an important aspect in stream data mining. Two distinct approaches are commonly employed for evaluating the quality of imputed values. First is the classification method, which can be used whenever the imputation task involves categorical or binary data as in Khan, Shahidul Islam, and Abu sayed Md Latiful Hoque (2020). To present a summary of the predictions made by a model on a set of data, comparing them with the actual true values, a confusion matrix is used.

Table 7: Confusion matrix

	Actual Positive	Actual negative
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

The top five classification methods used to assess imputation of missing values, as indicated in tables 4 and 6, include accuracy, recall, precision, specificity, and F1-measure. Accuracy is most used since it represents the overall correctness of the model and the ratio of correctly predicted instances to the total. If the primary goal of the imputation task is to recover as many true positive (correct imputed values) then recall would be a good option. FPR and FNR are two types of errors, where FPR is type 1 error, used to measure the proportion of actual negative instances that were incorrectly imputed as positive. It indicates the rate of incorrect imputations for negative values. On the other hand, FNR is type 2 error, it measures the proportion of actual positive instances that were incorrectly imputed as negative. In other words, it quantifies the rate of missing values that were not successfully imputed as positive. The choice between these methods depends on the goal of the study and the nature of the data. In addition, whenever there is imbalance in the datasets, some of these metrics may not be useful. Matthew's correlation coefficient (MCC) solves this issue, since it is less affected by imbalanced datasets compared to accuracy. Not only it considers all the component of the confusion matrix, but also balances between both precision and recall (Teh, Hui Yie, Andreas W. Kempa-Liehr, and Kevin I-Kai Wang. (2020)). We believe that the inclusion of MCC in future studies evaluating imputation techniques would contribute to a more thorough understanding of imputation quality, promoting advancements and progress in this crucial area of research.

The second approach for evaluating the quality is the regression metrics also called sample-wise discrepancy, throughout the selected studies, a variety of metrics have been employed to measure the effectiveness of fault correction or missing data imputation methods. They measure the imputation quality based on discrepancies between real and imputed values on a sample-by-sample basis (Shadbahr, Tolou, et al. (2022)). Among these, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) stand out as prominent indicators. RMSE, utilized 18 times in the 38 studies, measure the square root of the average squared errors between predicted errors and true values. Its interpretability is enhanced as it shares the same units as the vertical axis, providing a more intuitive understanding of the performance of the model. MAE, appearing 12 times, calculates the average of absolute errors, making it less sensitive to large differences compared to MSE or RMSE. MRE, used 5 times, quantifies the average relative errors between predictions and true values. MSE, used 4 times, measures the average squared errors (Teh, Hui Yie, Andreas W. Kempa-Liehr, and Kevin I-Kai Wang. (2020)), and NRMSE, employed 4 times, represents the normalized version of RMSE.

Given that the goal of imputation is not only to find the exact value of each missing value, but also to recover the correct distribution. To achieve such goal, Shadbahr, Tolou, et al. (2022) suggests that researchers should supplement the regression methods assessing imputation quality by using feature-wise distribution discrepancy, which evaluate how accurate the distribution individual features are reconstructed after imputation. The proposed metrics are the 2-Wasserstein distance, which calculates the distance between two probability distributions using optimal transport, the KS statistic, which evaluates differences between one-dimensional probability distributions, and the KL divergence, which approximates the true distribution of missing values. This method emphasizes the significance of not just imputing accurate values but also capturing the underlying distribution of the missing features for a more thorough evaluation of imputation quality. However, if the type of data is multi-dimensional a method called Sliced Wasserstein distance may be used for a better evaluation. It involves considering the joint distribution of multiple features simultaneously, allowing it to capture more nuanced differences that might be overlooked by methods that only consider marginal distributions (Shadbahr, Tolou, et al. (2022)).

4.3 Limitations

One limitation of our systematic review is the temporal constraint applied to the inclusion of studies, covering only the period from 2019 to 2023. This decision was made to focus on the most recent advancements in missing values and imputation techniques. However, it is acknowledged that this temporal restriction may have excluded valuable insights from earlier studies, which could have provided a historical context and a more comprehensive understanding of the evolution of imputation methods. Additionally, our study selection was primarily conducted through searches on Google Scholar and ArXiv. While these databases are rich sources of scholarly articles, we recognize that utilizing other literature collections, such as the IEEE Digital Library or PubMed (given the medical focus of several studies), could have yielded additional relevant studies.

5 Conclusion

In conclusion, our systematic review has provided a thorough exploration of novel imputation techniques as described in the included studies, focusing on the comparisons made and the methodologies used for handling missing values. We have also indicated the metrics used to assess the imputation methods, contributing to a broader understanding of their efficacy. we propose that future studies introducing novel imputation techniques should use an approach that considers both real-world and synthetic datasets. This approach ensures a strong evaluation that reflects the intricacies of real-world scenarios while providing controlled conditions for experimentation. Additionally, it is important to mention the type of missing data whether they are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR).

Furthermore, in the assessment of imputation methods, we recommend a strategy that includes both regression and classification metrics. This dual evaluation ensures better a better result in the evaluation. Also, we recommend researchers to incorporate methods that calculate the distribution of imputed missing values, to increase the depth and accuracy of the evaluation process. Future studies in missing values imputation can further develop this important field and promote more precise and adaptable imputation methods by taking these factors into account.

Author Contributions

AK, KM, SB, HA, and NA formed the keywords list, performed the literature search, gathered, and analyzed the studies, and wrote the systematic review. CB reviewed the paper.

Supplementary material

Both the included studies for RQ1 and RQ2 and the code used can be found in the following links

AHK011/Online-Imputation-Techniques-and-Quality-Assessment-for-Missing-Values-in-Data-Streams (github.com)

https://cloud.ovgu.de/s/WaPP4A3AZMFezFT

References:

1. Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27(1), e1873.
2. Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., & Kenward, M. G. (2023). Multiple imputation and its application. John Wiley & Sons.
3. Chen, K., Liang, X., Ma, Z., & Zhang, Z. (2022). GEDI: A graph-based end-to-end data imputation framework. arXiv preprint arXiv:2208.06573.
4. Cheng, C. H., Chan, C. P., & Sheu, Y. J. (2019). A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Engineering Applications of Artificial Intelligence, 81, 283-299.
5. Dai, Z., Bu, Z., & Long, Q. (2023, April). Multiple imputation with neural network Gaussian process for high-dimensional incomplete data. In Asian Conference on Machine Learning (pp. 265-279). PMLR.
6. Dong, W., Gao, S., Yang, X. et al. An Exploration of Online Missing Value Imputation in Non-stationary Data Stream. SN COMPUT. SCI. 2, 57 (2021). https://doi.org/10.1007/s42979-021-00459-1
7. Garcia, C., Leite, D., & Škrjanc, I. (2019). Incremental missing-data imputation for evolving fuzzy granular prediction. IEEE transactions on fuzzy systems, 28(10), 2348-2362.

8. Hamzah, F. B., Hamzah, F. M., Razali, S. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civil Engineering Journal, 7(9), 1608-1619.
9. Hamzah, F. B., Hamzah, F. M., Razali, S. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civil Engineering Journal, 7(9), 1608-1619.
10. Karmitsa, N., Taheri, S., Bagirov, A., & Mäkinen, P. (2020). Missing value imputation via clusterwise linear regression. IEEE Transactions on Knowledge and Data Engineering, 34(4), 1889-1901.
11. Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. Journal of big Data, 7(1), 1-21.
12. Khan, S.I., Hoque, A.S.M.L. SICE: an improved missing data imputation technique. J Big Data 7, 37 (2020). https://doi.org/10.1186/s40537-020-00313-w
13. Kim, S., Kim, H., Yun, E., Lee, H., Lee, J., & Lee, J. (2023). Probabilistic Imputation for Time-series Classification with Missing Data.
14. Kim, T., Ko, W., & Kim, J. (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. Applied Sciences, 9(1), 204.
15. Krempl, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., ... & Stefanowski, J. (2014). Open challenges for data stream mining research. *ACM SIGKDD explorations newsletter*, *16*(1), 1-10.
16. Kunicki, R., & Grzenda, M. (2021). Towards Increasing Open Data Adoption Through Stream Data Integration and Imputation. In Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34 (pp. 15-27). Springer International Publishing.
17. Lalande, F., & Doya, K. (2023). Numerical Data Imputation for Multimodal Data Sets: A Probabilistic Nearest-Neighbor Kernel Density Approach. arXiv preprint arXiv:2306.16906.
18. Lim, D. K., Rashid, N. U., Oliva, J. B., & Ibrahim, J. G. (2021). Unsupervised Imputation of Non-ignorably Missing Data Using Importance-Weighted Autoencoders. arXiv preprint arXiv:2101.07357.
19. Lin, Y., & Mehrotra, S. (2022). QUIP: Query-driven Missing Value Imputation. arXiv preprint arXiv:2204.00108.
20. Liu, S. H., Chrysanthopoulou, S. A., Chang, Q., Hunnicutt, J. N., & Lapane, K. L. (2019). Missing data in marginal structural models: a plasmode simulation study comparing multiple imputation and inverse probability weighting. Medical care, 57(3), 237.
21. Lo, A. W., Siah, K. W., & Wong, C. H. Machine Learning with Statistical Imputation for Predicting Drug Approvals, revised May 2019. Available at SSRN 2973611.
22. Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology, 110, 63-73.
23. Mera-Gaona, M., Neumann, U., Vargas-Canas, R., & López, D. M. (2021). Evaluating the impact of multivariate imputation by MICE in feature selection. Plos one, 16(7), e0254720.
24. 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.
25. Petrazzini, B.O., Naya, H., Lopez-Bello, F. et al. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Mining 14, 44 (2021). https://doi.org/10.1186/s13040-021-00274-7
26. Pourshahrokhi, N., Kouchaki, S., Kober, K. M., Miaskowski, C., & Barnaghi, P. (2021). A Hamiltonian Monte Carlo model for imputation and augmentation of healthcare data. arXiv preprint arXiv:2103.02349.
27. Razavi-Far, R., Saif, M., Palade, V., & Chakrabarti, S. (2022). An integrated framework for diagnosing process faults with incomplete features. Knowledge and Information Systems, 1-19.
28. Riggi, S., Riggi, D., & Riggi, F. (2020). Handling missing data in a neural network approach for the identification of charged particles in a multilayer detector. arXiv preprint arXiv:2004.05374.
29. Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., & Möller, M. (2019). Evaluating the accuracy of imputation methods in a five-way admixed population. Front Genet 10: 34.
30. Seth, N. (2021). Part 3: Topic modeling and Latent Dirichlet allocation (LDA) using Gensim and Sklearn. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/part-3-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/

31. Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., ... & Schönlieb, C. B. (2022). Classification of datasets with imputed missing values: Does imputation quality matter?. arXiv preprint arXiv:2206.08478.
32. Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., ... & Xiao, J. (2019). Comprehensive assessment of genotype imputation performance. Human Heredity, 83(3), 107-116.
33. Shu, Xiaokui, and Ron Cohen. "Natural Language Toolkit (NLTK)." (2010).
34. Sohrabi, C., Franchi, T., Mathew, G., Kerwan, A., Nicola, M., Griffin, M., ... & Agha, R. (2021). PRISMA 2020 statement: What's new and the importance of reporting guidelines. International Journal of Surgery, 88, 105918.
35. Spinelli, I., Scardapane, S., & Uncini, A. (2020). Missing data imputation with adversarially-trained graph convolutional networks. Neural Networks, 129, 249-260.
36. Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: A systematic review. Journal of Big Data, 7(1), 1-49.
37. Vazifehdan, M., Moattar, M. H., & Jalali, M. (2019). A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. Journal of King Saud University-Computer and Information Sciences, 31(2), 175-184.
38. Wang, D., Yan, Y., Qiu, R., Zhu, Y., Guan, K., Margenot, A., & Tong, H. (2023, August). Networked time series imputation via position-aware graph enhanced variational autoencoders. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 2256-2268).
39. Xu, X., Lai, T., Jahan, S., & Farid, F. (2022). Water and Sediment Analyse Using Predictive Models. arXiv preprint arXiv:2203.03422.
40. Xue, Y., Klabjan, D., & Luo, Y. (2019, December). Mixture-based multiple imputation model for clinical data with a temporal dimension. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 245-252). IEEE.
41. Zhang, X., Li, M., Wang, Y., & Fei, H. (2023). AmGCL: Feature Imputation of Attribute Missing Graph via Self-supervised Contrastive Learning. arXiv preprint arXiv:2305.03741.
42. Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Generation Computer Systems, 128, 63-72.
43. Zhao, J., Nie, Y., Ni, S., & Sun, X. (2020). Traffic data imputation and prediction: An efficient realization of deep learning. IEEE Access, 8, 46713-46722.
44. Zhao, Y., Landgrebe, E., Shekhtman, E., & Udell, M. (2022, June). Online missing value imputation and change point detection with the gaussian copula. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 8, pp. 9199-9207).
45. Zhu, X., Yang, J., Zhang, C., & Zhang, S. (2019). Efficient utilization of missing data in cost-sensitive learning. IEEE Transactions on Knowledge and Data Engineering, 33(6), 2425-2436.

.