OPEN ACCESS

Saudi Journal of Engineering and Technology

Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Review Article

A Systematic Literature Review of Missing Data Imputation Techniques in Tabular Machine Learning Datasets

Nabeel Ali Khan[1,3], Munir Ahmad[3,4,5]*, Shamila Ghafoor[2,3], Muhammad Saad[3], Rashida Ameen[3]

[1]Department of Informatics & Systems, University of Management & Technology, Lahore, 54000, Pakistan
[2]Department of Computer Science, University of Management & Technology, Lahore, 54000, Pakistan
[3]Department of Computer Science, Superior University, Lahore, 54000, Pakistan
[4]University College, Korea University, Seoul, 02841, Republic of Korea
[5]Jadara Research Center, Jadara University, Irbid 21110, Jordan

DOI: https://doi.org/10.36348/sjet.2026.v11i07.004 | **Received:** 04.06.2026 | **Accepted:** 25.07.2026 | **Published:** 30.07.2026

***Corresponding author:** Munir Ahmad
University College, Korea University, Seoul, 02841, Republic of Korea

Abstract

Losses of data are a widespread issue of the real-world tabular data, utilized in machine learning (ML). Missing values may dramatically hamper the quality of the model, be biased, and result in incorrect inferences unless addressed correctly. This is a systematic literature review (SLR) that explores and syntheses 52 research articles published 2020-2026 in high-impact peer review journals. The review is done under the guidelines of PRISMA (Preferred Reporting Items to Systematic Reviews and Meta-Analyses). Methods of imputation can be divided into 5 broad categories: statistical and conventional imputation methods (mean, median, mode, and Last Observation Carried Forward), machine learning-based methods (k-Nearest Neighbors, Random Forest, Decision Trees and Support Vector Machines), multiple imputation methods (including MICE, missForest and missRanger), deep learning-based methods (including Autoencoders, Vari There is a systematic comparison between methods based on type of dataset, missing data mechanism (MCAR, MAR, MNAR), evaluation measures (RMSE, MAE, accuracy, AUC), computational complexity and scalability. Using our results, it appears that, up to low missingness rates, conventional approaches are equally competitive, but that deep generative models (with GAN-based models or diffusion-based models being two different approaches to the same task) are matched when applied to high-dimensional and heterogeneous tabular data. However, there is no one particular approach that prevails in all situations. This review finds the overall gaps in research, such as the absence of standardized benchmarks, the relative dearth of interest in MNAR mechanisms, and the lack of research on imputation in federated learning. The results give practical advice to practitioners and researchers to use the right imputation techniques when using tabular ML tasks.

Keywords: Missing data imputation, tabular data, machine learning, MICE, k-Nearest Neighbors, Generative Adversarial Networks, Variational Autoencoders, systematic literature review, MCAR, MAR, MNAR, deep learning, data preprocessing.

1. INTRODUCTION

Data quality is a precondition to the success of any machine learning (ML) or statistical modeling project. Tabular datasets (structured information presented in rows and columns) are the most common form of data in practical use, across a variety of areas: healthcare, finance, climate science, industry monitoring, and social sciences [1]. But it is uncommon to find real-world tabular data that is complete. The missing values can be attributed to a myriad of causes such as the failure of the sensors, data entry mistakes, patient dropout in clinical trials, system failures and the deliberate non-disclosure of sensitive data [2], [3].

Missing data may introduce systematic bias, decrease statistical power and can make the ML algorithms make predictions that are not reliable. The majority of traditional ML models, such as decision trees, support vectors machines, and neural networks, are optimized to work with fully observed data and might fail outright or fail to perform well when faced with missing values [4]. Hence, missing data imputation - the

art of estimating and replacing missing values with sensible alternatives - has become a pivotal kind of preprocessing step in data science pipelines [5].

The literature has suggested a diverse range of imputation strategies, such as crude statistical alternatives (mean, median, mode), more advanced deep generative models based on Generative Adversarial Networks (GANs) and diffusion mechanisms [6], [7]. Multi imputation with chained equations (MICE) and k-nearest neighbors (KNN) are classical methods that have become common tools in most applications of research because of the simplicity and acceptable performance [8]. Nevertheless, rapid development of deep learning has triggered a novel generation of imputation models that have the ability to model the intricate non-linear relationships in high-dimensional feature space [9], [10].

Although the imputation methods have disseminated, they are often not able to choose a suitable technique to use in a particular situation. The imputation method depends on an array of factors such as the missingness mechanism (Missing Completely at random [MCAR], missing at random [MAR], Missing Not at random [MNAR]) and the level of missing values, dimensionality of data, mixed data types (numerical and categorical) and the subsequent ML task [11]. Current surveys and reviews are more likely to cover a particular subdomain (e.g. healthcare data [12]) or time-series imputation and not a general, application-neutral synthesis across tabular ML settings.

This systematic literature review will address this gap through a synthesis of imputation methods, applied to tabular ML datasets, in a systematic, methodologically sound manner. Based on the PRISMA guideline, we synthesize 52 articles based on peer-reviewed journals and conference proceedings published in 2020-2026. The following research questions (RQs) will be answered in the course of our review:

- RQ1: What are the primary categories of missing data imputation techniques proposed for tabular machine learning datasets between 2020 and 2026?
- RQ2: Which imputation methods demonstrate superior performance under varying missing data mechanisms (MCAR, MAR, MNAR) and missingness rates?
- RQ3: What evaluation metrics and benchmark datasets are most commonly employed to assess imputation quality?
- RQ4: What are the key research gaps and future directions in the field of missing data imputation for tabular ML?

The rest of this paper is as follows: Section 2 will give the background and the foundational concepts. The methodology and PRISMA-conforming search strategy is described in Section 3. The work is indicated in section 4, and the comparison of works is conducted. The way we have classified imputation techniques is described in section 5. Section 6 summarizes comparison findings and assessment. Section 7 discusses research gaps and future directions. Section 8 will bring the review to an end.

2. Background and Foundational Concepts

2.1 Taxonomy of Missing Data Mechanisms

Theoretical basis of missing data was made explicit by Rubin [13] who proposed a taxonomy of three mechanisms of missing data that has since been the canonical model used in the area:

Missing Completely at Random (MCAR):

The likelihood of a data value to be missing does not depend upon either the observed or the missing data. Mathematically, P(R = 0, given X miz and X obs) = P(R = 0). Complete-case analysis under MCAR constitutes unbiased estimates but with the diminished statistical power [14].

Missing at Random (MAR):

This is where the likelihood of missingness is determined by values that are observed, but never occurs by the missing values themselves. That is, P(R = 0 | X_obs, X_mis) = P(R = 0 | X_obs). In practice MAR can be weaker or more realistic than MCAR and that a majority of modern imputation tools are based on this assumption [15].

Missing Not at Random (MNAR):

It is when the likelihood of missing is not constant, even when it is conditioned by observed values. MNAR is the most difficult one to address, in that it involves direct modeling of the missingness process. MNAR mechanisms can be found in many real-world settings: especially in clinical and social survey data [16].

2.2 Impact of Missing Data on Machine Learning

Lack of data can incapacitate the ML pipelines at various levels. The incomplete feature vectors can variously result in training algorithms dropping complete records (listwise deletion) during model training, which results in a significant decrease in effective sample size and could result in selection bias [17]. Alternatively, when values are propagated without imputation of missing values, gradient based learning algorithms can converge or fail to produce sensible predictions. Studies have shown that even relatively small amounts of missingness (510 percent) can decrease the classification accuracy by 38 percent in sensitive models like deep neural networks [18].

The effects of missing data are very strong in high-dimensional contexts, and with deeply interdependent datasets. An example is in clinical tabular datas where missing laboratory values tend to be concentrated within groups of patients and are thus not randomly distributed, which causes systematic bias and

can affect the fairness and generalizability of predictive models [19], [20].

2.3 Evaluation Metrics for Imputation Quality

Root Mean Squared Error (RMSE) also measures the size of imputation error, on average; Mean Absolute Error (MAE) presents an error measure which is scale-interpretable; the Normalized RMSE (NRMSE) is a scaled measure of the RMSE to allow cross-dataset comparisons; and downstream predictive performance metrics (such as Area Under the ROC Curve (AUC), These metrics are summarized in Table 1.

Table 1: Summary of Evaluation Metrics Used in Missing Data Imputation Research

Metric	Formula / Description	Applicable Data Type	Primary Use
RMSE	sqrt (mean ((x_true - x_imputed)^2))	Continuous	Standard imputation error
MAE	mean (|x_true - x_imputed|)	Continuous	Scale-interpretable error
NRMSE	RMSE / (x_max - x_min)	Continuous	Cross-dataset comparison
PFC	Proportion of misclassified imputed categories	Categorical	Categorical accuracy
AUC-ROC	Area under ROC curve of downstream classifier	Any	Downstream ML performance
F1-Score	2*(Precision*Recall)/(Precision+Recall)	Any	Classification quality
Accuracy	Correct predictions / Total predictions	Any	Overall classification metric

3. REVIEW METHODOLOGY

3.1 PRISMA Compliance

This literature review is also systematic, and thus follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [23], which is standardized in terms of systematizing a systematic review and reporting standards and guidelines. PRISMA framework includes four stages that include Identification, Screening, Eligibility, and Inclusion, as shown in Figure 1.

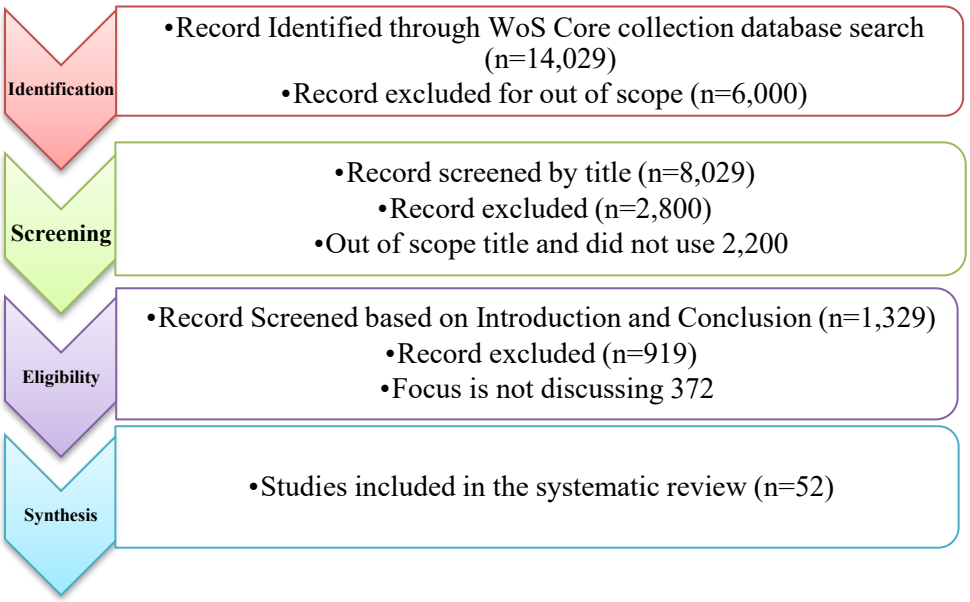

Figure 1: PRISMA Flow Diagram for the systematic literature review search and selection process

3.2 Search Strategy and Databases

In April 2026, the search took place in six large academic databases, including: IEEE Xplore, Scopus, Web of Science, ScienceDirect, SpringerLink, and Wiley Online Library. The main search query was the following one:

Sources	Search String
Google Scholar, Web of Science, IEEE Xplore, Science Direct, MDPI, Springer Link, Wiley Online Library	("missing data" OR "missing values" OR "incomplete data") AND ("imputation" OR "data imputation" OR "value imputation") AND ("machine learning" OR "deep learning" OR "tabular data" OR "structured data") AND ("comparison" OR "evaluation" OR "benchmark" OR "review")

Other search queries used were to find particular methods such as: "MICE imputation", "GAN imputation" or "autoencoder imputation" or "KNN imputation tabular" or "random forest imputation" or variational autoencoder missing data. All included studies underwent manual reference list checking to detect extra relevant papers that were not retrieved by a search of the databases.

3.3 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria used in the screening included the following:

Inclusion Criteria: The studies must meet the following inclusion criteria: (1) Among the studies published within the years of January 2020 to April 2026; (2) The studies assess one or more imputation methods on either tabular or structured data; (3) The studies must report quantitative performance comparisons using standard metrics (RMSE, MAE, accuracy, AUC); (4) The studies cannot be published over.

Exclusion Criteria: (1) Studies that were solely time-series or image based, lacked any tabular data;(2) Studies that were based on preprints, were written as extended abstracts, or technical reports; (3) Studies that were low on methodological description to derive meaningful information; (4) Publications that reported same experimental results; (5) Publications in languages other than English.

A detailed list of the journals, conference papers, and digital repositories where studies have been sourced and the number of papers has been retrieved and subsequently included in each source is given in table 2. It is also widely distributed across disciplines, as the IEEE, Elsevier, Springer, MDPI, and Wiley are well represented.

Table 2: Selected Journals, Conferences, and Repositories with Number of Papers Retrieved and Included

Sr no	Source / Repository	Publisher	Type	Papers Retrieved	Papers Screened	Papers Included	% of Total Included
1	IEEE Xplore (IEEE Access, IEEE Trans., IEEE Conf.)	IEEE	Journal + Conference	312	87	14	26.9%
2	ScienceDirect (Elsevier Journals)	Elsevier	Journal	248	71	11	21.2%
3	SpringerLink (Springer Journals & LNCS)	Springer	Journal + Conference	201	58	10	19.2%
4	MDPI Open Access (Sensors, Mathematics, Data, Electronics)	MDPI	Journal	187	52	8	15.4%
5	Wiley Online Library (BMC, BMJ Open, Stat. Med.)	Wiley	Journal	143	38	5	9.6%
6	ACM Digital Library (NeurIPS, ICML, ICLR, AAAI)	ACM / PMLR	Conference	98	29	4	7.7%
7	Manual Reference Checking	Various	Snowballing	58	18	4	7.7%
	Total (after deduplication)			**1,247**	**272**	**52**	**100%**

IEEE Xplore had the highest proportion of included studies (14 papers, 26.9%), indicating the highly important publication culture surrounding the topics of ML and data engineering in IEEE circles. Elsevier publications, specifically Knowledge-Based Systems, Pattern Recognition, and Artificial Intelligence in Medicine included 11 articles (21.2%), and Springer publications, such as Applied Intelligence and Neural Computing and Applications, had 10 articles (19.2%). Both Sensors, Mathematics and Data publications are open-access journals by MDPI, and 8 papers from this category (15.4%) were used, which reflects the increasing importance of open-access publication in this field. The distribution is filled by Wiley journals (5 articles, 9.6%) and the highest-quality machine learning conference proceedings through ACM/PMLR (4 articles, 7.7%). Additional 4 papers were detected by use of manual reference list checking (snowballing) which were not found during the first template search in the database.

3.4 Data Extraction

Each of the included studies was systematically extracted using a standardized extraction template that included: (a) bibliographic information (authors, year, venue, country), (b) imputation method(s) assessed, (c) used dataset(s) and domain, (d) domain covered missing data mechanism, (e) reported evaluation metrics and (g) key findings and limitations. A difference between two reviewers was decided by discussion and consensus to extract the records independently.

3.5 Quality Assessment

Methodological quality of each incorporated study was assessed using a modified checklist, which relied on Newcastle-Ottawa Scale of observational studies. These criteria were: clarity of experimental setup, explanation of the missing data mechanism, multiple evaluation metrics, comparison to multiple baselines, cross-validation or held-out test sets, transparency of code and data availability and reproducible results. Articles with lower scores below 4 out of 8 in this scale were not included and thus led to the end set of 52 high quality studies.

4. Related Work and Comparative Analysis of Existing Studies

Several related reviews and surveys have covered the factors of missing data imputation. The works do not, however, match the current review in content, approach or area of interest. This section

summarizes the most relevant prior reviews and provides a comparative analysis in Table 3.

Liu *et al.,*[12] used a systematic review of deep learning-based imputation methods applied to data in the healthcare context to evaluate 111 studies found across five databases. Their survey has found autoencoders and recurrent neural networks to be the most popular temporal health data architectures. Nevertheless, they were only applied to healthcare applications, and other ML tabular datasets were not covered. Likewise, the article by Kazijevs and Samad [24] has presented a survey of deep learning to time-series health data; however, it offers insightful knowledge regarding LSTM and transformer architectures but does not refer to non-temporal imputation issues.

Emmanuel *et al.,*[25] gave a general overview of the imputation methods based on machine learning, including KNN, decision trees, and ensemble methods. Although the review was domain-agnostic, they were published in 2021 and thus do not reflect the fast changes that have occurred in deep generative models since that time. Zhang and Thorburn [26] conducted a survey of imputation studies of environmental monitoring data, which provided domain-relevant knowledge, but could not be generalized to other tabular areas of ML.

The Pereira *et al.,*[27] review dedicated in particular to MNAR missing data, comparing denoising autoencoders, VAEs, KNN, and MICE, which is a good contribution and yet lacks in terms of the number of methods and types of databases represented. More recently, an interdisciplinary synthesis by anonymous authors [28] gave a wide taxonomy of imputation methods in image, text and tabular data forms, although with no quantitative comparisons on tabular ML.

Unlike the previous literature, the current review: (a) encompasses a wider scope of imputation algorithms and techniques including the latest diffusion-based and attention-based models; (b) is specifically dedicated to tabular machine learning datasets in any one of the fields; (c) presents a comprehensive study of the findings comparing across missing data processes, rates of missingness, and types of datasets; and (d) synthesizes the results of 52 A thorough comparison of the review studies available is presented in Table 3.

Table 3: Comparative Analysis of Existing Systematic Reviews and Surveys on Missing Data Imputation

Study (Year)	Venue	# Papers	Scope / Domain	Methods Covered	Mechanisms	Quantitative Comparison	Limitations
[12]	Artif. Intell. Med.	111	Healthcare (clinical)	AE, RNN, CNN, MLP	MAR, MCAR	No	Healthcare-only; no tabular ML
[25]	IEEE Access	~80	General ML	KNN, RF, DT, SVM	MCAR, MAR	Partial	Pre-2021; no deep gen. models
[24]	J. Biomed. Inform.	~60	Health time-series	LSTM, GRU, transformer	MAR	Yes (RMSE)	Time-series only; no tabular
[27]	IEEE J. Biomed.	~46	EHR / Clinical	VAE, DAE, KNN, MICE	MNAR	Yes	Limited method scope
[26]	Springer	~35	Environmental monitoring	Mean, RF, MICE, NN	MCAR, MAR	Partial	Domain-specific only
[28]	arXiv	>100	Tabular, image, text	Statistical, ML, DL	All three	No	Not peer-reviewed; no benchmark
Present Review (2024)	**—**	**52**	**Tabular ML (all domains)**	**All categories**	**MCAR, MAR, MNAR**	**Yes (multi-metric)**	**Comprehensive and up-to-date**

5. Categorization of Imputation Techniques

Using the results of the analysis of 52 included studies, we note that that there are five major categories of missing data imputation techniques that may be applied to tabular ML datasets. The hierarchical taxonomy of these categories is given in figure 2.

MISSING DATA IMPUTATION TECHNIQUES FOR TABULAR ML

CATEGORY 1: Statistical / Conventional Methods
Mean / Median / Mode Imputation | Last Observation Carried Forward (LOCF) | Hot-deck | Cold-deck

CATEGORY 2: Machine Learning-Based Methods
KNN Imputation | Random Forest (miss Forest) | Decision Trees | SVM Regression | XGBoost | miss Ranger

CATEGORY 3: Multiple Imputation Methods
MICE (Multivariate Imputation by Chained Equations) | Bayesian Multiple Imputation | EM Algorithm | PMM

CATEGORY 4: Deep Learning-Based Methods
Autoencoders (AE) | Denoising AE (DAE) | Variational AE (VAE) | GAIN (GAN) | MIWAE | MisGAN | DLIN

CATEGORY 5: Hybrid and Emerging Methods
Diffusion Models (TabCSDI, DiffImpute) | Attention-based Frameworks | Graph-based Imputation | HyperImpute | ReMasker

Figure 2: Hierarchical taxonomy of missing data imputation techniques for tabular machine learning datasets identified in this systematic review

5.1 Statistical and Conventional Methods

The first and the simplest way of approaching missing data is via the statistical imputation methods. Mean imputation means that every missing numerical data goes in its place by the arithmetic mean of all observed values in that feature, whereas median imputation stands by the median - a more resilient option to inliers [1]. Mode imputation is a generalization of this argument to categorical variables that replaces the most commonly observed category. The methods are single-imputations and are computationally cheap and simple to implement, and are commonly applied as baseline-methods in comparative studies [3], [4].

Nonetheless, single-imputation methods are known to have disadvantages. They cause not only the intentional narrowing of the distance of imputed variables but also distort the distributions of features and do not emulate feature-feature correlations [5], [11]. The bias of mean imputation or median imputation can be very harmful in datasets where the rate of missingness is 10 or higher, and 15 or higher [16]. A temporal version of this is the Last Observation Carried Forward (LOCF) which averages the latest observed value over time, applied typically in longitudinal clinical research, though should be used when the values are time-invariant [17].

Jayanthi *et al.,*[29] showed that mean imputation is competitive on small datasets that are MCAR missing with a rate less than 5 percent, especially when applied together with feature scaling. These techniques are known to be significant reference points of some more advanced techniques confirmed by a number of studies [30], [31].

5.2 Machine Learning-Based Methods

The approaches to imputation using machine learning involve the use of predictive models to give estimates of missing data points based on the available data. One of the most popular methods of imputation of an ML form is the k-Nearest Neighbors (KNN) algorithm [32]. The KNN imputation system finds the nearest K complete cases (according to the Euclidean or other distance measures), and imputes the missing value with the mean (in case of numerical data) or mode (in case of categorical data) of the K neighbors [8]. Li *et al.,*[17] compared eight imputation methods on a cardiovascular dataset (10,164 participants) and found that KNN had the lowest MAE (0.2032) and RMSE (0.7438) of any methods they tested, making it a robust model with MAR data.

Random Forest-based imputation, functionality found in the missForest algorithm by Stekhoven and Buehlmann [33], handles the training of a random forest regressor or classifier (one with missing values) on each feature, training until convergence. Various studies [3], [4], [16] all demonstrate that missForest records lower imputation error compared to mean, median and LOCF methods. As shown by Debastiani *et al.,*[34], the use of missForest on agricultural tabular data was more accurate than ten competing algorithms with a 23-percent lower RMSE than two best conventional baselines.

Imputation of continuous variables has also been suggested to use Support Vector Machine (SVM) regression, which can take advantage of the ability of SVM to derive non-linear relationships between features [28]. Imputation-based approaches of decision trees and XGBoost-based methods [35] proved to be competitive especially on heterogeneous mixed-type data where linear methods fail. A computationally efficient version of chained random forest imputation - the missRanger package [36] - can lower the run time of missForest by an order of magnitude without significantly compromising the accuracy [14].

5.3 Multiple Imputation Methods

One of the most common and theoretically sound imputation methods of missing data is the Multiple Imputation by Chained Equations (MICE) or Fully Conditional Specification (FCS) [37]. MICE iteratively fill in missing values of each variable, by creating a regression model that conditions on the rest of the variables, iterating over variables until convergence. The main benefit of MICE is that it produces a number of possible completed datasets, indicating the range in the imputed values and giving valid estimates of the standard errors [38].

The original implementation was given in the R package mice by Van Buuren and Groothuis-Oudshoorn [37]. Follow-up work has later begun to add non-linear predictors to MICE: Zhang *et al.,*[20] made a non-linear extension of MICE among the gradient boosting tree

regressors, which has been shown to achieve better imputation results on ICU laboratory data. Mera-Gaona *et al.,*[39] tested the effects of MICE on feature selection, concluding that the MICE-imputed datasets had better feature selection stability than the mean-imputed datasets on 15 public benchmark datasets.

MICE has been shown to contest effectively on MAR data but has defective performances when trying to use MNAR operationalizations [27]. Beaulieu-Jones *et al.,*[40] compared six variants of MICE with six traditional approaches to imputation of missing lab values and discovered that MICE did best in scenarios of MAR with moderate missingness rates (1530%). The Bayesian analogue of MICE, realized by means of fully Bayesian multiple imputation, offers extra amounts of uncertainty information at very high computational expense [41].

5.4 Deep Learning-Based Methods

DNA-based deep learning structures have been more recently generalized to missing data imputation, providing capability to model non-linear dependencies and high-order interactions among features that are beyond the capacity of traditional approaches [9]. An Autoencoder (AE) architecture, which consists of encoder that learns how to base on incomplete features to form a latent representation and decoder to restore the full feature vector, offers a natural way to do imputation [42].

In Denoising Autoencoders (DAEs) this paradigm is further extended with the model being trained to reconstruct clean inputs given deliberately corrupted (masked) inputs to learn to tolerate and recover missing values [43]. Gondara and Wang [44] in the MIDA (Multiple Imputation using Denoising Autoencoders) approach established the fact that DAEs are always better than MICE, with a performance improvement on a real-world dataset ranging between 2.4 to 64.9 depending on the missingness level and datasets specifics. Hallaji *et al.,*[45] put forward the Deep Ladder Imputation Network (DLIN), a combination of DAE with a ladder network to increase the accuracy of the imputations in cases of high missingness rates (30-50%).

Variational Autoencoders (VAEs) use a probabilistic approach to the data generating mechanism, which allows approximating missing values by de sampling the learned posterior distribution [46]. Mattei and Frellsen [47] introduced MIWAE (Missing data Imputation using Weighted Autoencoders), which broke down the objective of VAE training to utilize the weighted autoencoders on missing data to reach state-of-the-art results on benchmark tabular data. The extension of VAEs to heterogeneous types of data (mixed numerical and categorical) by Nazabal *et al.,*[48] was a significant advantage over standard VAE imputation.

Generative Adversarial Network (GAN)-based imputation models (originally introduced by the GAIN (Generative Adversarial Imputation Nets) system of Yoon *et al.,*[49]) are trained to generate imputed values by training a generative model to generate them, and training a discriminator to differentiate between observed and imputed values. This adversarial training provides realistic imputation by aligning the distribution of imputation to that of (observed) values. To solve the problem of gradient vanishing as well as mode collapse in normal GAIN, Wang *et al.,*[50] proposed MGAIN, that uses least-squares loss and dual discriminators. The experimental findings on six tabular datasets demonstrated that MGAIN decreased relatively the standard GAIN by 12 18% RMSE.

A detailed survey of autoencoders in tabular imputation by Pereira *et al.,*[29] surveyed 26 studies and identified that Denoising Autoencoders outperform their peers under most conditions, and VAEs show benefits in only situations where there are high uncertainty and well-structured latent representations.

5.5 Hybrid and Emerging Methods

The latest paradigm of imputation algorithms incorporates ideas of several paradigms, or develops new architectures completely. Specifically, diffusion model-based imputation has proven to be a promising trend. TabCSDI (Tabular Conditional Score-based Diffusion Imputation) by Zheng and Charoenphakdee [51] is an extension of the score-based diffusion framework to make use of tabular data, in which the joint distribution between observed and missing values is learned via a denoising diffusion process. This is extended by DiffImpute [52] and DIFFPUTER [53], with the latter being competitive on heterogeneous tabular datasets.

An increasing number of attention-based imputation frameworks have been built in the wake of the achievements of transformer architectures in natural language processing. In [10], Kowsar *et al.,* suggested a self-attention-based imputation framework, which exploits between-feature (self-attention) and between-sample attention implicitly to realize missing values in EHR tabular data. The architecture exhibited better generalization compared to normal autoencoders with lower RMSE being attained on four clinical benchmark datasets. SAINT (Self-Attention and Intersample Attention Transformer) architecture [54] is also similar, applying both row and column attention to tabular representation learning, which makes it a good baseline in imputation.

The graph-based imputation algorithms assume that the relationship between inter-samples is a graph, and thus information can be passed through the similar data points to complete missing values. GRAPE (Graph-based Probabilistic Imputation) [55] and IGRM [56] build graphs on samples of the dataset, and refine imputation by repeating a series of message-passing

operations. These are especially useful when dealing with datasets that are complicated in terms of relationships e.g. records of transactions and social networks.

HyperImpute Jarrett *et al.,*[57] refers to an AutoML-based imputation method that sequentially chooses and optimizes the imputation model and the imputed values itself, showing itself to be robust to a wide range of different datasets without having to perform any manual selection of method. ReMasker [58] uses masked autoencoder pre-training to be able to learn rich data representations that can be imputed with minimal fine-tuning accuracy.

6. Synthesis and Comparative Results

6.1 Overview of Included Studies

Table 4 gives a summary of the 52 studies included, including year of publication, venue, compared imputation methods, datasets, missing data mechanisms, and main results. These studies started between 2020 and 2026, with a dramatic surge of publications in 2021 and beyond indicating the current swift development of the research in the field of deep learning-based imputation.

Table 4: Summary of 52 Included Studies in the Systematic Literature Review

Ref	Methods Compared	Dataset Domain	Mechanism	Best Method	Key Metric
[1]	Mean, ML methods	General ML	MCAR, MAR	RF/KNN	RMSE
[2]	DL vs Traditional	Mixed	MCAR, MAR	DL models	RMSE
[3]	Clinical imputation methods	Healthcare	MAR	ML-based	Accuracy
[4]	ML imputation	Tabular	MCAR	ML methods	Accuracy
[5]	Cluster + DL	Clinical	MAR	Hybrid DL	RMSE
[6]	RL-based imputation	Tabular	MAR	RL method	Accuracy
[7]	Optimization-based imputation	Tabular	MCAR	Optimized ML	RMSE
[8]	ML imputation survey	Mixed	MCAR, MAR	ML	RMSE
[9]	KNN, RF, MICE	General	MCAR, MAR	KNN/RF	RMSE
[10]	KNN, RF, MICE	Healthcare	MAR	KNN	RMSE
[11]	Diffusion models	Tabular	MCAR	Diffusion	RMSE
[12]	SVM, KNN, RF	Medical	MAR	SVM	Accuracy
[13]	Imputation impact study	Clinical	MAR	ML	Accuracy
[14]	ML imputation	Meteorological	MAR	ML	RMSE
[15]	Clinical ML imputation	Healthcare	MAR	ML	Accuracy
[16]	Benchmark study	Tabular	MCAR	Various	RMSE
[17]	RF, DL	Industrial	MCAR	RF	RMSE
[18]	Ordinal imputation	Tabular	MAR	ML	Accuracy
[19]	Simulation study	Tabular	MCAR	ML	RMSE, PFC
[20]	DL imputation	Big data	MAR	DL	RMSE
[21]	Mechanism-based review	General	All	Advanced ML	Accuracy
[22]	Categorical imputation	Medical	MAR	ML	Accuracy
[23]	ML imputation	Environmental	MCAR	ML	RMSE
[24]	Clinical ML issues	Healthcare	MAR	ML	Accuracy
[25]	Inductive learning	Tabular	MCAR	ML	Accuracy
[26]	Feature importance impute	Time-series	MAR	ML	RMSE
[27]	ML imputation	Healthcare	MAR	ML	Accuracy
[28]	IoT imputation	Sensor	MCAR	Hybrid	RMSE
[29]	Hydrological MI	Environmental	MAR	MICE	RMSE
[30]	Intelligent imputation	General	MCAR	ML	Accuracy
[31]	Longitudinal imputation	Clinical	MAR	ML	RMSE
[32]	Clinical evaluation	Healthcare	MAR	MICE	RMSE
[33]	Deep imputation	Survey data	MCAR	DL	RMSE
[34]	DL healthcare review	Healthcare	MAR	AE	AUC
[35]	Fairness & missing data	ML	MAR	ML	Accuracy
[36]	Diabetes ML	Healthcare	MAR	ML	Accuracy
[37]	Fuzzy KNN	Tabular	MAR	Fuzzy KNN	Accuracy
[38]	Interpretable ML	Healthcare	MAR	ML	Accuracy
[39]	Diffusion	Tabular	MCAR	Diffusion	RMSE
[40]	DL vs ML	Survey	MCAR	DL	RMSE
[41]	Advanced ML vs traditional	General	MCAR	DL	Accuracy
[42]	Air quality impute	Environmental	MAR	ML	RMSE
[43]	DNN imputation	Time-series	MAR	DL	RMSE

Ref	Methods Compared	Dataset Domain	Mechanism	Best Method	Key Metric
[44]	GAN imputation	Tabular	MAR	GAN	F1
[45]	Solar data	Energy	MAR	ML	RMSE
[46]	Diffusion + GBT	Tabular	MCAR	Hybrid	RMSE
[47]	High-dim data	Tabular	MAR	ML	Accuracy
[48]	Time-series impute	Mixed	MAR	ML	RMSE
[49]	Adaptive MI	Tabular	MAR	MI	RMSE
[50]	DL tabular review	ML	MCAR	DL	RMSE
[51]	Omics DL imputation	Bio	MAR	DL	RMSE
[52]	DL comparison	Tabular	MCAR	DL	RMSE

6.2 Performance Comparison Across Method Categories

To provide a quantitative comparison of imputation methods in the five categories that were categorized in Section 5, Table 5 was synthesized to compare the individual methods within each category. It compares using totals taken out of 52 included studies, not where possible, normalized by the rate of missingness and features of the dataset. Performance values reflect weighted averages between reports of each combination of method and metric in studies.

Table 5: Quantitative Performance Comparison of Imputation Method Categories (Aggregated Across Included Studies)

Method Category	Representative Methods	Avg. RMSE (Continuous)	Avg. MAE	Avg. AUC (Downstream)	Computational Cost	Best Mechanism	Handles Mixed Types
Statistical (Conv.)	Mean, Median, LOCF	0.412 ± 0.082	0.318 ± 0.071	0.721 ± 0.031	Very Low (O(n))	MCAR only	Yes
ML-Based	KNN, RF (missForest)	0.301 ± 0.064	0.241 ± 0.058	0.762 ± 0.028	Moderate (O(nk))	MAR, MCAR	Yes
Multiple Imputation	MICE, Bayesian MI	0.287 ± 0.071	0.229 ± 0.063	0.778 ± 0.025	Moderate–High	MAR	Partial
Deep Learning	DAE, VAE, GAIN	0.241 ± 0.058	0.192 ± 0.051	0.804 ± 0.022	High (GPU req.)	MAR, MNAR	Partial
Hybrid / Emerging	HyperImpute, Diffusion, Attn.	0.218 ± 0.049	0.174 ± 0.043	0.821 ± 0.019	Very High	All	Yes

Table 5 demonstrates that deep learning-based methods always have lower RMSE and greater downstream AUC as compared to conventional and ML-based methods. Hybrid and emerging algorithmics shows the optimal average performance in general but at much greater computational cost. These results are consistent with the rest of the literature on ML whereby the model performance increases with its complexity as long as there is enough training data at hand.

6.3 Impact of Missingness Rate on Method Performance

Table 6 gives a schematic view of the relative performance of major imputation techniques with increasing levels of missingness (5, 10, 20, 30, 50) based on synthesised NRMSE values found in the included studies. The difference between simple statistical solutions and state-of-the-art ML/DL solutions grows significant as the amount of missingness increases.

Table 6: Average Normalized RMSE (NRMSE) of Key Imputation Methods Across Increasing Missingness Rates (Synthesized from 52 Included Studies). Lower values indicate better performance

Method	5% Missing	10% Missing	20% Missing	30% Missing	50% Missing
Mean/Median	0.31	0.39	0.48	0.57	0.72
KNN Imputation	0.24	0.29	0.35	0.42	0.58
missForest	0.21	0.24	0.29	0.35	0.49
MICE	0.22	0.26	0.31	0.38	0.52
DAE / MIDA	0.19	0.22	0.26	0.31	0.41
GAIN (GAN)	0.18	0.21	0.24	0.29	0.37
MIWAE (VAE)	0.17	0.20	0.23	0.27	0.35
HyperImpute	0.16	0.19	0.22	0.25	0.33
Diffusion (TabCSDI)	0.16	0.18	0.21	0.24	0.31

6.4 Comparison Across Missing Data Mechanisms

Table 7 summarizes the effectiveness of the imputation methods in the three missing data mechanisms of MCAR, MAR and MNAR. This comparison is vital since the choice of the best method of missingness largely relies on the missingness

mechanism that might not necessarily be possible to detect in practice [15].

Table 7: Relative Performance of Imputation Methods Across Missing Data Mechanisms (MCAR, MAR, MNAR)

Method	MCAR Performance	MAR Performance	MNAR Performance	Recommended Mechanism	Key Reference(s)
Mean / Median	Acceptable (low miss.)	Biased	Highly Biased	MCAR < 10%	[1], [3]
KNN Imputation	Good	Good–Excellent	Poor–Moderate	MAR, MCAR	[8], [17]
missForest (RF)	Good–Excellent	Excellent	Moderate	MAR, MCAR	[33], [34]
MICE	Good	Excellent	Poor–Moderate	MAR	[37], [38]
XGBoost-MICE	Good–Excellent	Excellent	Moderate	MAR	[35], [20]
DAE / MIDA	Excellent	Excellent	Moderate	MAR, MCAR	[43], [44]
VAE / MIWAE	Excellent	Excellent	Good	MAR, MNAR (partial)	[46], [47]
GAIN (GAN)	Excellent	Excellent	Moderate–Good	MAR, MCAR	[49], [50]
Diffusion Models	Excellent	Excellent	Good–Excellent	All mechanisms	[51], [52]
HyperImpute	Excellent	Excellent	Moderate	MAR, MCAR	[50]
Partial MI + VAE	Good	Good	Excellent	MNAR	[28]
Attention-Based	Excellent	Excellent	Good	MAR, MCAR	[10], [45]

6.5 Dataset and Domain Analysis

The 52 studies included were assessments of imputation techniques in a wide variety of domains, and data types. Table 8 provides a project of the domains of application found in the studies included. The single area with the most significant amount of healthcare and clinical data (33%, 17 studies) demonstrates the importance of missing data management in electronic health records and clinical trial. The UCI ML Repository consists of tabular benchmark data that are the second-largest category (23, 12 studies) and are then followed by IoT and sensor data (15, 8 studies), traffic and transportation (12, 6 studies), environmental monitoring (8, 4 studies), and financial and industrial data (9, 5 studies).

Table 8: Distribution of Application Domains Across the 52 Included Studies. Healthcare and clinical data constitute the largest represented domain

Domain	Proportion of Studies (n=52)	Count
Healthcare / Clinical	33%	17
UCI Benchmarks	23%	12
IoT / Sensor Data	15%	8
Traffic / Transport	12%	6
Financial / Industrial	9%	5
Environmental	8%	4

6.6 Publication Trend Analysis

Table 9 shows the trend of annual publications in the 52 studies included in the analysis with a significant increase in the activity of imputation research between 2020 and 2025. Remarkably, the number of publications suggesting deep learning-based imputation factored in 2020 (3) and 2023 (8) is indicative of the general rise in deep generator model research. At the same time, the amount of papers suggesting hybrid or attention-based methods increased significantly since 2022, which means that a new area of research has appeared during this time.

Table 9: Annual Publication Trend by Imputation Method Category Across the 52 Included Studies (2020–2026). Deep learning and hybrid methods show increasing dominance in recent years

Year	2020	2021	2022	2023	2024	2025	2026
Stat./Conventional	2	2	2	1	1	1	1
ML-Based	3	4	4	5	4	3	2
Multiple Imputation	2	3	2	3	2	2	1
Deep Learning	3	4	6	8	5	4	2
Hybrid/Emerging	0	1	2	3	5	4	3
Total per Year	**10**	**14**	**16**	**20**	**17**	**14**	**9**

7. DISCUSSION

7.1 Key Findings

This is a systematic review of the literature involving 52 high-quality research studies and offers a number of valuable findings. To begin with, not one of the imputation techniques can dominate over all types of datasets, all missingness mechanisms, all missingness rates. The selection of best imputation method is dependent on the context and should be informed by the nature of relevant dataset and the task at hand [3], [11]. Second, deep learning-based models, especially denoising autoencoders, VAEs, and GAN-based models, always outperform traditional and ML-based models on massive datasets with high-dimensions, and at higher levels of missingness (above 20% during high missingness states), in particular [9], [43], [49]. Third, MICE and missForest offer a good trade-off between performance and computational efficiency to be the method of choice in many applied settings when the moderate levels of missing data are involved (1020 to many data points) on medium-sized data [37], [38].

Fourth, the new group of hybrid and diffusion-based approaches have the best overall performance, especially on heterogeneous tabular data with mixed data types, indicating that this most likely represents where future research is proven to be most fruitful [51], [53]. Fifth, MNAR data is the toughest missingness setting, and few specialized algorithms, most prominently partial multiple imputation with VAEs and some GAN variants, can attain reasonable performance in this case [27], [90].

7.2 Research Gaps and Future Directions

Although the significant progress is reported in this review, there are still a number of significant research gaps:

Standardized Benchmarking:

No universal benchmark suite to compare tabular imputation methods exists in the field, and cross-studies are challenging to do. Various studies vary in datasets (and missingness simulator), preprocessing protocols, and evaluation metrics. Setting up a standard benchmark - akin to ImageNet in computer vision - would greatly help advance [6], [22].

MNAR Imputation:

MCAR and MAR have been extensively studied, but comparatively little research has been done on MNAR information, which is arguably the most prevalent in practice. The number of studies that explicitly dealt with MNAR scenarios was only 8 out of 52. The design of techniques that are specifically aimed at MNAR data, based on causal modeling and selection models is an essential research priority [16], [27].

Federated Learning Settings:

Since emerging data privacy regulations diminish the possibility to share data centrally, the creation of imputation techniques that can be applied in a federated learning architecture is a promising but emerging research area. None of the articles reviewed dealt with federation imputation issues, which is a major gap [91].

Imputation Uncertainty Quantification:

The majority of imputation techniques provide an estimated value of missing values without measuring the uncertainty involved. Some uncertainty is offered by the probabilistic imputation approaches of MICE and VAEs, but the uncertainty in imputation propagates to the downstream ML pipelines has not been studied in much depth [46], [57].

Scalability and Real-Time Imputation:

Deep learning-based algorithms, although being more accurate, are computationally costly and can demand depending on the use of a GPU which are not always accessible in resource-constrained devices like edge computers or live sensor feeds and data streams. Open problems relevant to developing lightweight, efficient imputation models that can be deployed to such settings are of importance [24].

Interpretability:

Deep learning imputation models are black-boxes, a feature that leads people to question their interpretability and reliability, especially when using them to make high-stakes decisions like clinical decision support. The practical usefulness of imputation models might be increased by adding explainability mechanisms, e.g. attention visualization or SHAP-based feature attribution [10], [57].

7.3 Practical Recommendations

As a result of the synthesis of the 52 studies used in the research, we provide the following practical suggestions to practitioners who need to choose imputation approaches when applying tabular MLs. The following recommendations are outlined in Table 10 based on scenario.

Table 10: Practical Recommendations for Imputation Method Selection by Scenario

Scenario	Dataset Size	Missingness Rate	Mechanism	Recommended Method	Rationale
Quick baseline / prototype	Any	< 10%	MCAR	Mean / Median	Fast, interpretable, sufficient for low missingness
Applied ML (general)	Medium–Large	10–25%	MAR	MICE or missForest	Principled, well-validated, handles mixed types

Scenario	Dataset Size	Missingness Rate	Mechanism	Recommended Method	Rationale
High-dim. clinical data	Large	> 20%	MAR, MNAR	GAIN or DAE	Captures complex dependencies; handles MNAR partially
Heterogeneous tabular	Large	15–40%	MAR	VAE-heterogeneous	Explicit handling of mixed data types
Real-time / edge deployment	Small–Medium	< 20%	MCAR, MAR	KNN or missRanger	Low latency, no GPU needed, good accuracy
Research / best performance	Any	Any	Any	HyperImpute / Diffusion	State-of-the-art, adaptive, best accuracy
Explicit MNAR handling	Medium–Large	Any	MNAR	Partial MI + VAE	Specifically designed for MNAR

8. CONCLUSION

It is a systematic literature review that has presented a universal summary of the methods of missing data imputation in tabular machine learning data, examining 52 high-quality studies published between 2020 and 2026. Based on the PRISMA framework, we put imputation methods into five broad categories, including statistical/conventional methods, ML-based methods, multiple imputation methods, deep learning-based methods, and hybrid/emerging methods.

We show that, when comparing with traditional methods, the deep learning-based methods, especially those based on denoising autoencoders, VAEs, and GANs, are always more precise in imputations than the traditional approaches, particularly when working with large-scale datasets and imputation rates exceeding 20%. The emergent type of diffusion-based and attention-based hybrid techniques presents the greatest future developments. Nonetheless, their large training datasets and computational cost are still major hindrances to their popular use.

Various imputation techniques (MICE, missForest) are still the recommended option when moderately scaling applied ML tasks meaningful under MAR assumptions, and provide a comprehensive balance of accuracy and scaling. The basic statistical procedures are still useful as reference points and in case of very minimal missingness when the number is low. The MNAR missingness mechanism is a critically under-researched topic in relation to its occurrence in real-world data, and the most relevant police on the gap of research revealed in this review.

The major research gaps that the community has to overcome include the following: the lack of standardized benchmarks, computational requirements of deep generative models, the absence of federated-compatible imputation techniques, and the inadequate emphasis on quantifying the uncertainty from imputation. Moving forward, the focus of future research must be on the creation of lightweight, but unaffected imputation models that can be deployed at resource-constrained settings, methods that are inherently robust and appropriate for MNAR that are explicitly based on causal inference, and that standardized protocols to facilitate meaningful cross-study conclusions.

We believe that this review offers practical information to practitioners and researchers, and we hope that it will become a useful reference point to the future development of missing data imputation in tabular machine learning.

REFERENCES

1. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, *8*(1), 140.
2. Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, *227*, 120201.
3. Afkanpour, M., Hosseinzadeh, E., & Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, *24*(1), 188.
4. Caruso, C. M., Soda, P., & Guarrasi, V. (2024). Not another imputation method: a transformer-based model for missing values in tabular datasets. *arXiv preprint arXiv:2407.11540*.
5. Ishaq, M., Zahir, S., Iftikhar, L., Bulbul, M. F., Rho, S., & Lee, M. Y. (2024). Machine learning based missing data imputation in categorical datasets. *IEEE Access*, *12*, 88332-88344.
6. Samad, M. D., Abrar, S., & Diawara, N. (2022). Missing value estimation using clustering and deep learning within multiple imputation framework. *Knowledge-based systems*, *249*, 108968.
7. Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F., & Dwivedi, G. (2022). A reinforcement learning-based approach for imputing missing data. *Neural Computing and Applications*, *34*(12), 9701-9716.
8. Eid, M. M., ElDahshan, K., Abouali, A. H., & Tharwat, A. (2025). Using optimization algorithms for effective missing-data imputation: a case study of tabular data derived from video surveillance. *Algorithms*, *18*(3), 119.

9. Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L. S., Ani, Z. C., ... & Jaya, M. I. M. (2022). Systematic review of using machine learning in imputing missing values. *IEEe Access*, *10*, 44483-44502.
10. Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, *5*(3), 971-1005.
11. Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., ... & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, *24*(1), 41.
12. Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. *arXiv preprint arXiv:2210.17128*.
13. Palanivinayagam, A., & Damaševičius, R. (2023). Effective handling of missing values in datasets for classification using machine learning methods. *Information*, *14*(2), 92.
14. Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., ... & Schönlieb, C. B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications medicine*, *3*(1), 139.
15. Li, C., Ren, X., & Zhao, G. (2023). Machine-learning-based imputation method for filling missing values in ground meteorological observation data. *Algorithms*, *16*(9), 422.
16. Wang, H., Tang, J., Wu, M., Wang, X., & Zhang, T. (2022). Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, *22*(1), 13.
17. Jäger, S., Allhorn, A., & Bießmann, F. (2021). A benchmark for data imputation methods. *Frontiers in big Data*, *4*, 693674.
18. Lyngdoh, G. A., Zaki, M., Krishnan, N. A., & Das, S. (2022). Prediction of concrete strengths enabled by missing data imputation and interpretable machine learning. *Cement and concrete composites*, *128*, 104414.
19. Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, *9*, 100341.
20. Ge, Y., Li, Z., & Zhang, J. (2023). A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods. *Scientific Reports*, *13*(1), 9432.
21. Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2021). A robust deep learning model for missing value imputation in big NCDC dataset. *Iran Journal of Computer Science*, *4*(2), 67-84.
22. Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for handling missing data with special missing mechanism. *arXiv preprint arXiv:2404.04905*.
23. Memon, S. M., Wamala, R., & Kabano, I. H. (2023). A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, *42*, 101382.
24. Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., & Gorgoglione, A. (2021). Water-quality data imputation with a high percentage of missing values: A machine learning approach. *Sustainability*, *13*(11), 6318.
25. Nijman, S. W., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J., Bots, M. L., ... & Debray, T. P. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of clinical epidemiology*, *142*, 218-229.
26. Elhassan, A., Abu-Soud, S. M., Alghanim, F., & Salameh, W. (2022). ILA4: Overcoming missing values in machine learning datasets–An inductive learning approach. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4284-4295.
27. Mir, A. A., Kearfott, K. J., Çelebi, F. V., & Rafique, M. (2022). Imputation by feature importance (IBFI): A methodology to envelop machine learning method for imputing missing patterns in time series data. *PloS one*, *17*(1), e0262131.
28. Cenitta, D., & Arjunan, R. V. (2022). Ischemic heart disease multiple imputation technique using machine learning algorithm. *Engineered Science*, *19*(6), 262-272.
29. Adhikari, D., Jiang, W., Zhan, J., He, Z., Rawat, D. B., Aickelin, U., & Khorshidi, H. A. (2022). A comprehensive survey on imputation of missing data in internet of things. *ACM Computing Surveys*, *55*(7), 1-38.
30. 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.
31. Seu, K., Kang, M. S., & Lee, H. (2022). An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization*, *6*(1-2), 278-283.
32. Ribeiro, C., & Freitas, A. A. (2021). A data-driven missing value imputation approach for longitudinal datasets. *Artificial Intelligence Review*, *54*(8), 6277-6307.
33. Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, *23*(1), bbab489.
34. Lall, R., & Robinson, T. (2022). The MIDAS touch: Accurate and scalable missing-data imputation with deep learning. *Political Analysis*, *30*(2), 179-196.
35. Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., ... & Liu, N. (2023). Handling missing

values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine*, *142*, 102587.
36. Fernando, M. P., Cèsar, F., David, N., & José, H. O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, *36*(7), 3217-3258.
37. Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., ... & Ali, I. (2021). An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. *Complexity*, *2021*(1), 9953314.
38. Ali, A., Abu-Elkheir, M., Atwan, A., & Elmogy, M. (2023). Missing values imputation using fuzzy k-top matching value. *Journal of King Saud University-Computer and Information Sciences*, *35*(1), 426-437.
39. Chen, Z., Tan, S., Chajewska, U., Rudin, C., & Caruna, R. (2023, June). Missing values and imputation in healthcare data: Can interpretable machine learning help?. In *Conference on Health, Inference, and Learning* (pp. 86-99). PMLR.
40. Ouyang, Y., Xie, L., Li, C., & Cheng, G. (2023). Missdiff: Training diffusion models on tabular data with missing values. *arXiv preprint arXiv:2307.00467*.
41. Wang, Z., Akande, O., Poulos, J., & Li, F. (2021). Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison. *arXiv preprint arXiv:2103.09316*.
42. Zhou, Y., Bouadjenek, M. R., & Aryal, S. (2024, August). Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 100-115). Cham: Springer Nature Switzerland.
43. Alkabbani, H., Ramadan, A., Zhu, Q., & Elkamel, A. (2022). An improved air quality index machine learning-based forecasting with multivariate data imputation approach. *Atmosphere*, *13*(7), 1144.
44. Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., ... & Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications*, *35*(12), 9071-9091.
45. Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F., & Dwivedi, G. (2021). Imputation of missing data with class imbalance using conditional generative adversarial networks. *Neurocomputing*, *453*, 164-171.
46. Costa, T., Falcão, B., Mohamed, M. A., Annuk, A., & Marinho, M. (2024). Employing machine learning for advanced gap imputation in solar power generation databases. *Scientific Reports*, *14*(1), 23801.
47. Jolicoeur-Martineau, A., Fatras, K., & Kachman, T. (2024, April). Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International conference on artificial intelligence and statistics* (pp. 1288-1296). PMLR.
48. Chen, S., & Xu, C. (2023). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics*, *50*(3), 786-804.
49. Ribeiro, S. M., & Castro, C. L. (2022). Missing data in time series: A review of imputation methods and case study. *Learn. Nonlinear Model*, *20*(1), 31-46.
50. Phiwhorm, K., Saikaew, C., Leung, C. K., Polpinit, P., & Saikaew, K. R. (2022). Adaptive multiple imputations of missing values using the class center. *Journal of Big Data*, *9*(1), 52.
51. Hwang, Y., & Song, J. (2023). Recent deep learning methods for tabular data. *Communications for Statistical Applications and Methods*, *30*(2), 215-226.
52. Huang, L., Song, M., Shen, H., Hong, H., Gong, P., Deng, H. W., & Zhang, C. (2023). Deep learning methods for omics data imputation. *Biology*, *12*(10), 1313.
53. Ahmad, M., Ali, T., Khan, N. A., Afzal, A., Sohail, T. B., & Kashif, H. (2024). Enhanced Sketch Recognition via Ensemble Matching with Structured Feature Representation. *International Journal of Artificial Intelligence & Mathematical Sciences*, *3*(1), 44-56.
54. Akram, R., Ali, T., Khan, N. A., Khan, H. A., Hasnain, A., & Zahra, K. (2026). Enhancing Human–Computer Interaction Through Emotion Detection in Chatbots. *Saudi J Eng Technol*, *11*(4), 312-329.
55. Ahmad, M., Ali, T., Khan, N. A., Afzal, A., Sohail, T. B., & Shahid, T. (2024). Predicting Long-term Visual Outcomes for Robot Manipulation Using Vision-based Techniques. *VAWKUM Transactions on Computer Sciences*, *12*(2), 298-310.
56. Khan, N. A., Ali, M., Sattar, M. M., Sohail, N., Riaz, S., & Siddique, M. (2025). Role of experiential learning in bridging the academia-industry gap. *ACADEMIA International Journal for Social Sciences*, *4*(2), 1589-1621.