IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 11, 2022, accepted March 12, 2022, date of publication April 22, 2022, date of current version April 29, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160841

Systematic Review of Using Machine Learning in Imputing Missing Values

MUSTAFA ALABADLA[1], FATIMAH SIDI[1], (Member, IEEE), ISKANDAR ISHAK[1], HAMIDAH IBRAHIM[1], (Member, IEEE), LILLY SURIANI AFFENDEY[1], ZAFIENAS CHE ANI[1], MARZANAH A. JABAR[2], UMAR ALI BUKAR[2,3], NAVIN KUMAR DEVARAJ[4], AHMAD SOBRI MUDA[5], ANAS THAREK[5], NORITAH OMAR[6], AND M. IZHAM MOHD JAYA[7], (Member, IEEE)

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[2]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[3]Department of Mathematical Sciences, Computer Science Unit, Taraba State University, Jalingo 00234, Nigeria
[4]Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[5]Department of Radiology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[6]Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[7]Department of Software Engineering, Faculty of Computing, Universiti Malaysia Pahang (UMP), Pekan, Pahang 26600, Malaysia

Corresponding author: Fatimah Sidi (fatimah@upm.edu.my)

This work was supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme under Grant FRGS/1/2020/ICT06/UPM/02/1.

ABSTRACT Missing data are a universal data quality problem in many domains, leading to misleading analysis and inaccurate decisions. Much research has been done to investigate the different mechanisms of missing data and the proper techniques in handling various data types. In the last decade, machine learning has been utilized to replace conventional methods to address the problem of missing values more efficiently. By studying and analyzing recently proposed methods using machine learning approaches, vital adoptions in accuracy, performance, and time consumed can be highlighted. This study aimed to help data analysts and researchers address the limitations of machine learning imputation methods by conducting a systematic literature review to provide a comprehensive overview of using such methods to impute missing values. Novel proposed machine learning approaches used for data imputation are analyzed and summarized to assist researchers in selecting a proper machine learning method based on several factors and settings. The review was performed on research studies published between 2016 and 2021 on adopting machine learning to impute missing values, focusing on their strengths and limitations. A total of 684 research articles from various scientific databases were analyzed using search engines, and 94 of them were selected as primary studies. Finally, several recommendations were given to guide future researchers in applying machine learning to impute missing values.

INDEX TERMS Systematic literature review, data imputation, data mining, missingness, data preprocessing, data quality.

I. INTRODUCTION

Missing values are one of several data quality challenges that often occur in real-world datasets. This common issue usually affects data analytics performance by causing high bias and producing low accuracy. Missing values can happen for multiple reasons, such as respondents' refusal to answer, manual typing errors, or equipment malfunctions [1]–[3]. The occurrence of missing data is inevitable during the data collection stage. Therefore, datasets that include missing values should be treated before entering the preprocessing phase.

Moreover, Janssen *et al.* [4] highlight that having a complete dataset can greatly influence the decision-making process in an organization. For example, low-quality data will lead to inaccurate analysis, which will result in the wrong

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

decisions being made. Donald B. Rubin [5] proposes that missing values be categorized under three main mechanisms, each containing a different pattern. The first mechanism is missing completely at random (MCAR), whereby the missing values here have no relationship or dependence on observed, unobserved, or the missing data itself. The second mechanism is missing at random (MAR), in which the missing values have a relationship with the observed values. The third and last mechanism is missing not at random (MNAR). These mechanisms are only applicable if none of the previous mechanisms is valid, and the missing values are usually related to unobserved predictors or the missing value itself [6].

Several imputation approaches were proposed to handle this issue, starting from the most basic approaches, such as listwise and pairwise deletion methods. However, excluding some values from a dataset can greatly reduce its performance. Although simple and straightforward, methods such as mean substitution can produce a high bias if the percentages of the missing value are quite high [7]. More advanced and promising imputation methods, such as multiple imputations, provide a new way of dealing with missing values by creating parallel datasets and calculating the estimated values for missing values individually [8]. The main advantage of using multiple imputation methods over a single imputation is handling data uncertainty while using different imputation models in conjunction with it [9].

More focus has been given to utilizing machine learning (ML) techniques to overcome the missing data issue in the last decade. Multiple methods have been proposed using supervised, semisupervised and unsupervised ML techniques. These methods impute the missing values by dividing the dataset into training and test sets and learning from the observed variables. Normally, ML methods have some data type restrictions and cannot maintain the same performance while dealing with different values [10]. Thus, it is crucial to understand the data pattern before conducting data imputation to select the most suitable ML method [10].

Several studies have published the missing data problem in the ML domain [7], [11]–[14]. However, these studies have been dispersed among different journals and conference proceedings. Although there are various proposed ML models and methods to impute missing values in different domains, there is a lack in tracking the main body of research and the improvement in results according to reported studies. Hence, a clear overview is needed to illustrate how ML contributes to improving data quality and handling missing data issues. Accordingly, comprehensive insight can be gained by applying ML to impute missing data.

This article presents the findings of a systematic literature review (SLR) performed on ML approaches for missing value imputation, specifically focusing on the recently proposed ML methods, improvements in performance, and hybrid models. It is crucial to understand the missing value ratio and its mechanism before conducting imputations using ML techniques. The main objectives of this study are to determine recent trends of ML applications in handling missing values, evaluate the ML method used in addressing the missing data problem, identify the limitations and strengths in these methods, and determine possible research ideas and research improvements in future work.

The article is organized as follows. Related work is discussed in Section 2. Section 3 describes the research method. The results of the survey are presented in Section 4. Section 5 presents the research findings and future research directions. Finally, a conclusion is provided in Section 6.

TABLE 1. Related works summary.

Ref.	Summary
[13]	This study conducted a comprehensive literature review on single and multiple imputation approaches to resolving the missing values issue for air pollution datasets. Although the authors did not include ML methods, research gaps can still be seen, which will assist researchers in utilizing ML approaches to overcome the missing values and achieve higher performance.
[14]	The authors proposed a framework to assist researchers and analysis practitioners in selecting the right imputation method using several intertwined factors. In their study, four different ML approaches were analyzed in addition to two ad hoc methods. Their study's findings recommended using K-NN for the missing completely at random (MCAR) mechanism to achieve higher performance.
[15]	The authors review integrative imputation approaches in bioinformatics, focusing on deep learning methods to address missing data in multiomics datasets. Several factors that affect the accuracy of data imputation, such as missing values mechanism, missingness ratio, dataset size, and noise level, were identified by investigating these methods. Another challenge found in the study is the application of deep learning techniques such as autoencoder to address missing data in multiomics settings.
[16]	The authors investigated several imputation techniques in hydrology. The authors focused on model-based methods and ML techniques in addition to the deletion method and single imputation methods. One of the key findings in this research is that ML approaches are more flexible in handling missing values issues and can deal with higher data dimensions, which leads to better performance and accuracy. However, interpreting the results can be considered a major challenge in such cases

II. RELATED WORKS

In the last decade, the number of published studies in the area of data quality has increased dramatically. However, only a few have reviewed the literature on utilizing ML techniques to handle missing values. Table 1 summarizes the related literature review studies.

From the findings of the summarized related works in Table 1, we note the significant role that ML techniques play in improving the performance of handling missing values in different domains. Data imputation performance can be evaluated using different factors, such as accuracy, time consumed, and computational cost. The study conducted by Timur *et al.* [14] shows that using deep learning combined with multiple imputations by chained equations (MICE) for data imputation can lead to higher performance. The

proposed method was evaluated by measuring its accuracy, sensitivity, and specificity. Webb-Robertson *et al.* [15] evaluated several data imputation methods based on the root-mean-square error (RMSE) and classification accuracy. It was found that selecting the appropriate imputation method depends on the dataset itself and the analysis's purpose. The studies above compared ML methods with conventional data imputation methods, such as deletion and single imputation approaches. Thomas and Rajabi [7] provided a descriptive analysis of ML applications in imputing missing values related to experimental settings such as datatypes, missingness mechanisms, platforms used, missing ratios, and dataset characteristics used in several studies.

The studies referred to above are dispersed in different journals and conference proceedings. Unlike previous works, the main aim of this study is to assist data analysts and researchers in selecting the most suitable ML approach to improve data quality based on several factors and settings, such as the purpose of the study, dataset characteristics, domain problems, missing value mechanisms and ML techniques. The selection of an appropriate ML method can lead to a noticeable improvement in data imputation performance. Furthermore, it is crucial to understand the various types of ML techniques to reduce the time consumption required for imputing the missing values.

Moreover, this systematic review focuses on the limitations and strengths of recent ML applications in imputing missing values. Additionally, this study examines the evaluation methods and domains and provides several future research directions by proposing a taxonomy for selecting the best ML imputation method.

III. RESEARCH METHOD

The study follows the guidelines for conducting a systematic literature review proposed by Kitchenham *et al.* [16]. The execution process of this systematic literature review contains eight main stages, as shown in Figure 1. The first stage is identifying the research questions that will be answered by extracting relevant data from the selected primary studies. The second stage is the search strategy, which includes identifying suitable keywords to search for related studies in addition to selecting research resources from well-known science databases, journals, and conference proceedings in the related domain. The search process used the selected keywords and logical operators in research resources to identify relevant studies. In the third stage, inclusion and exclusion criteria were defined to select the most relevant studies. In the fourth stage, several quality criteria are defined to avoid bias and to ensure that the selected studies had enough information. In the fifth stage, collected research papers were filtered, and primary studies were selected based on their abstracts for relevancy and based on predetermined inclusion and exclusion criteria of the full text. The required data are extracted from primary studies and synthesized to answer the research questions in the sixth and seventh stages. Extracted data are analyzed and listed as data items

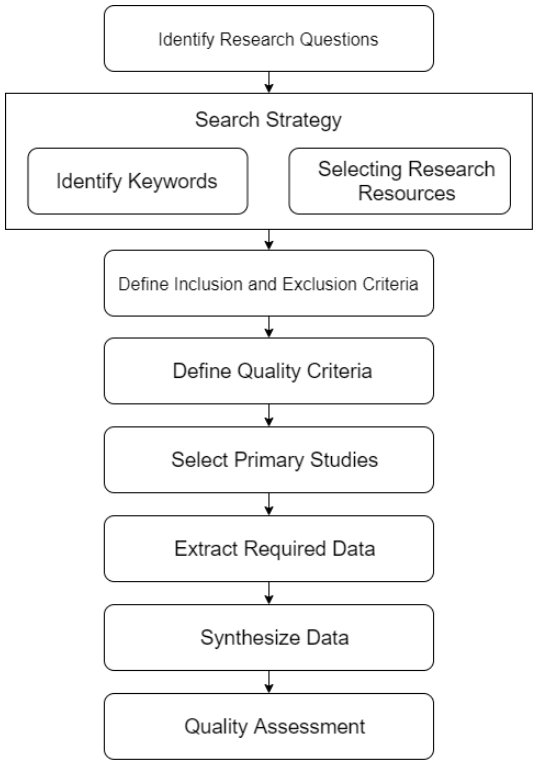

FIGURE 1. Systematic literature review execution process.

with categories and subcategories. Finally, the eighth stage involved evaluating primary studies against each quality criterion, and the score for every quality criterion was calculated.

A. RESEARCH QUESTIONS

The main objective of this article is to investigate and analyze novel ML techniques in addressing missing value issues by highlighting significant performance factors to achieve the best results. Therefore, to address these research gaps, this systematic literature review addresses the following research questions:

RQ. 1 What are the trends in machine learning imputation methods over the last five years?

RQ. 2 What are the most commonly used machine learning approaches to impute missing values?

RQ. 3 What are the characteristics of datasets and types of analysis used in the primary studies?

RQ. 4 What measurement factors are used to evaluate the ML imputation methods?

RQ. 5 What are the limitations and strengths of applying ML methods to impute missing values?

RQ1 aimed to provide an overview of recent trends in ML techniques used to impute missing values and to show how often ML methods are used instead of conventional imputation approaches. RQ2 was intended to categorize ML techniques used for data imputation and analyze the implementation process.

RQ3 sought to investigate the characteristics of the collected datasets and how ML imputation methods were adapted to handle different types of analysis. The answer to this question will analyze and reveal the commonly used dataset settings in primary studies using descriptive analysis to illustrate the diversity of different settings used for the imputed datasets.

RQ4 aimed to analyze the performance metrics used for evaluating the ML imputation methods. This question is intended to clarify the factors behind the selection process, whether it is prediction accuracy, random error, or execution time. The answer to this question will show what measurement factors are used the most to determine the level of adequacy in evaluating the proposed methods.

RQ5 was intended to highlight limitations and strong points in using ML techniques to impute missing values. Answering this question would provide a clear guideline for future work regarding the use of ML in data imputation.

B. SEARCH STRATEGY

The search strategy was developed based on two main factors: identifying keywords and selecting research resources. Data were collected using an automatic search feature in every research resource. The keywords used in the search process were selected based on two main categories: missing values and ML. The search targeted both the research title and the abstract. Table 2 illustrates the main categories with their respective keywords.

The reason for including ML keywords is to find studies that used ML techniques but did not mention the 'ML' keyword in their research title or abstract. Since several studies have used ML techniques for different purposes, we were able to identify specific keywords about missing values and gather data from related studies only.

The automatic search was conducted in several selected research databases. Five different research databases were selected, as shown in Table 3. These databases were chosen from a list of most used electronic databases for researchers based on a study done by Chen *et al.* in [17]. We selected five research databases from the list to reduce redundancy in the collected data. Another reason for selecting these databases is that they include the advanced search feature whereby logical operators can be used. In addition, the filtering option allowed us to limit the results based on the publication year.

C. INCLUSION AND EXCLUSION CRITERIA

The automatic search for relevant studies was conducted in selected research databases using the keywords listed in Table 2. The scope of the search process targeted articles published between 2016 and 2021. Two iterations were conducted on the automatic search results to select the primary studies. In the first iteration, the title of the research, in addition to the abstract, was scanned to select the related studies based on the SLR objectives. Additionally, any duplications in the research results were removed. In the second iteration, the whole text was scanned to double confirm whether the selected studies from the first iteration were aligned to the main objectives of this SLR study. Furthermore, we applied the following inclusion and exclusion criteria in this iteration to select the most relevant primary studies.

TABLE 2. Main categories and their respective keywords.

Index	Category	Keywords
KC01	Missing Values	'missing values', 'missing data', 'data missingness', 'missing data imputation', 'data imputation', 'imputing missing values', 'incomplete data.'
KC02	ML	'ML', 'supervised', 'unsupervised', 'semisupervised', 'classification', 'decision tree', 'k-nearest neighbor', 'naïve Bayes', 'random forest', 'clustering', 'neural network', 'deep learning'

TABLE 3. Selected research databases.

Index	Database Name	URL
RD01	SCOPUS	https://www.scopus.com
RD02	IEEE Xplore	https://ieeexplore.ieee.org
RD03	ACM Digital Library	https://dl.acm.org
RD04	ScienceDirect	https://www.sciencedirect.com
RD05	Wiley Online Library	https://onlinelibrary.wiley.com/

1) Inclusion criteria:
   a) The publication date of the study was between January 2016 and December 2021. The reason for selecting this period is because ML techniques became commonly used by researchers at the starting year to investigate state-of-the-art proposed ML methods.
   b) The article uses a single ML approach, ensemble, or hybridized model with other imputation methods.
   c) The article focuses on solving the missing values problem in a dataset.
   d) The proposed ML method is evaluated with other ML imputation methods.
   e) The article is published in the English language.
   f) The article is published in a journal or proceedings with peer review.
2) Exclusion criteria:
   a) The article should not be an abstract only or editorial.
   b) The article should not use conventional imputation methods instead of ML techniques.
   c) The article should not focus on image classification problems.
   d) The article's main aim should not be to improve any factor other than the data imputation performance.

TABLE 4. Search queries used in selected databases.

Database Name	Search Query
SCOPUS	TITLE-ABS-KEY((data AND imputation) AND (missing AND value*) AND ((machine AND learning) OR supervised OR unsupervised OR semisupervised)) PUBYEAR > 2015 PUBYEAR < 2022
IEEE Xplore	("missing values" AND "data imputation") AND (("machine learning ") OR "supervised" OR "unsupervised" OR "semisupervised")
ACM Digital Library	(("missing values") AND (("machine learning ") OR "supervised" OR "unsupervised" OR "semisupervised"))
ScienceDirect	"missing values" AND ("imputation") AND ("machine learning" OR "supervised" OR "unsupervised" OR "semisupervised")
Wiley Online Library	("missing values") AND (("machine learning ") OR "supervised" OR "unsupervised" OR "semisupervised")

TABLE 5. Data item extraction form.

Category	Subcategory	Research Question (RQ)
Reference Information	Title	RQ1
	Publish Year	
	Authors	
	Publication source	
Research Focus	Purpose	RQ1
	Methodology	
ML Technique	Proposed approach	RQ2
	Single or hybrid	
	ML type	
	Selection method	
Dataset	Number of datasets	RQ3
	Dataset sources	
	Data types	
	Missing values ratio	
	Missingness mechanism	
	Type of analysis	
Performance	Performance metrics	RQ4
Evaluation	Evaluation methods	
Findings	Limitations	RQ5
	Strength	

e) The article should not be focused on predicting a case instead of imputing missing values.

The inclusion criteria (a)-(f) were implemented during the first iteration, which included scanning the title of the research in addition to the abstract. On the other hand, exclusion criteria (a)-(e) were implemented during the second iteration of the primary study selection process, whereby the full texts of the selected studies were reviewed to filter out the nonrelated articles. Table 4 demonstrates the search queries used in each of the five selected research databases using keywords and logical operators in advanced search features.

D. QUALITY CRITERIA

The main aim of this section is to ensure that primary studies have enough information to answer the research questions. Each criterion is referred to as 'QAC', which stands for Quality Assessment Criteria. Studies are evaluated using the following quality assessment questions:

QAC.1 Does the paper use ML approaches for imputing missing values?

QAC.2 Were the research purpose and methodology explained clearly?

QAC.3 Is the proposed approach evaluated with other ML imputation methods?

QAC.4 Did the researchers explain the performance measurements used?

QAC.5 Does the paper cover the strengths and limitations of the proposed method?

E. SELECTION OF PRIMARY STUDIES

The automatic search on online databases for using ML methods in imputing the missing values returned 684 studies. After conducting the first iteration by scanning the title and abstract, nonrelated studies were excluded, and 168 studies were identified as related. In the second iteration, 74 studies were dropped after applying the inclusion and exclusion criteria. Hence, 94 primary studies were used to answer the research questions, as shown in Figure 2.

F. DATA EXTRACTION

The search process for related studies was restricted to designated electronic databases, journals, and conference proceedings. Every article was evaluated based on the categories and subcategories listed in Table 5. Listed categories were selected based on the research questions to provide answers to them. Data items were extracted from articles using automatic and manual search procedures.

The first category reference information was used for both documentation and to answer RQ1 by checking whether the

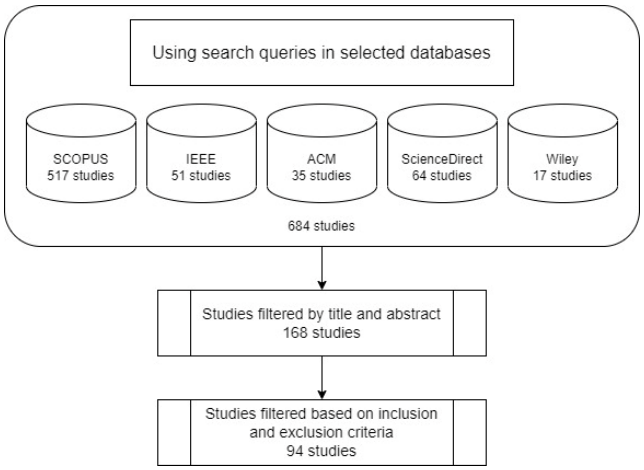

FIGURE 2. Primary study selection procedure.

paper was published within the last six years or not. On the other hand, the research focus category is used to identify the trends of ML methods used for data imputation to complete the answer for RQ1.

Next, the ML technique category was listed to address RQ2. RQ2 was intended to identify the most commonly used ML technique to impute missing values based on the information extracted from articles, such as the type of method used. Third, RQ3 intended to identify the dataset settings used for missing value imputation, including the number of datasets used in the experiments, sources of these datasets, data types imputed, and the percentage and mechanism of missingness. In addition, it investigates the type of analysis adopted to handle missing values in collected datasets.

The following is RQ4, which aims to determine the performance factors used to evaluate the proposed ML methods against other existing imputation methods in addition to evaluation methods that include the research design followed in the primary studies.

Finally, the findings answered RQ5 by determining the limitations and strengths of applying ML methods to impute missing values. Usually, this information is very important for several reasons. First, it can help researchers find research gaps and what is missing. Second, it shows the weak points that may be overcome using different approaches. Third, it may include some suggested solutions that can be considered in future work.

G. DATA SYNTHES

In this phase, extracted data items are fused together, in addition to the amount of data needed to answer each research question. The following explains the data synthesis approach used in this study: For RQ1, RQ2, RQ3, and RQ4, data items were aggregated and presented using a descriptive analysis technique (quantitative). Furthermore, a narrative synthesis approach was used to formulate the information per RQ5, which was obtained from different papers.

TABLE 6. Quality assessment results.

Criteria	Responding score	Total score
QAC01	{0, 0.5, 1} (No, Partially, yes)	94 studies (100%)
QAC02	{0, 0.5, 1} (No, Partially, yes)	94 studies (100%)
QAC03	{0, 0.5, 1} (No, Partially, yes)	86 studies (91%)
QAC04	{0, 0.5, 1} (No, Partially, yes)	78 studies (83%)
QAC05	{0, 0.5, 1} (No, Partially, yes)	57 studies (61%)

H. QUALITY ASSESSMENT

In addition to the inclusion and exclusion evaluation, each primary study was fully assessed using specific Quality Assessment Criteria (QAC) questions to avoid bias and increase the selection of literature [18]. The assessment is conducted by assigning a score between 0 and 1 for each primary study. If the study answers the QAC question, it is given a score of 1, and if it does not provide a full answer, it is given a score of 0.5. However, if the study fails to answer the QAC question, it is given a score of 0. The total score is calculated by summing all QAC question scores.

After conducting the quality assessment for each primary study, the total score of the selected primary studies was $>$ 60% against each QAC, as shown in Table 6. This finding indicates that the primary studies have sufficient information about imputing missing values using ML techniques.

IV. RESULT AND ANALYSIS

This systematic literature review was carried out based on the procedure explained in the previous section. The results of this study are shown by answering each of the proposed research questions using data extracted from the selected research database.

A. RQ1. WHAT ARE THE TRENDS IN ML IMPUTATION METHODS OVER THE LAST FIVE YEARS?

Using ML in imputing missing values has improved the overall performance of prediction and data analysis in comparison to conventional imputation methods [19]. Nkonyana *et al.* [20] indicated that non-ML methods might reduce sample size and that variability reduction produces high bias. With current advancements in technology, ML has the advantage of high computational resources and has proven its ability to overcome these issues by estimating the missing values with high accuracy and enhanced data analysis performance [21].

Figure 3 demonstrates the distribution of selected studies between 2016 and 2021. Although the number of studies remained steady in 2016 and 2017, there was a significant increase in the number of studies that utilized ML approaches. In particular, since 2016, the number of studies remained constant at nine studies per year until 2018, whereby a

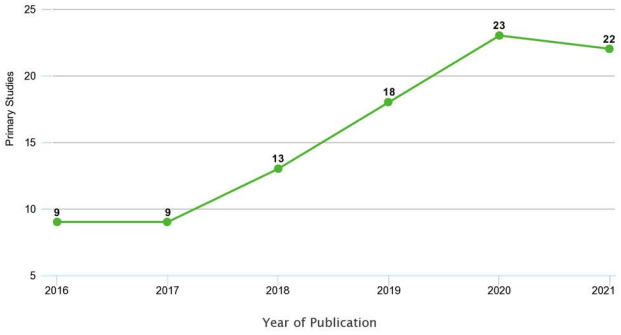

FIGURE 3. Distribution of primary studies over the last six years.

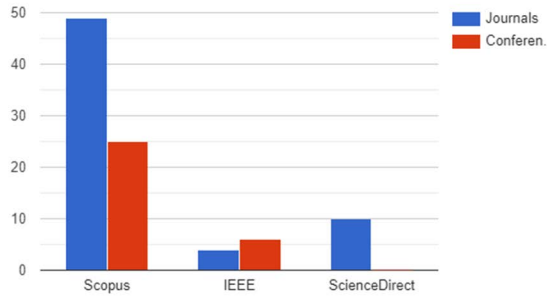

FIGURE 4. Publication trend of primary studies.

remarkable rising trend was noticed compared to previous years. This increase continues steadily until recently in 2020, which shows a growth of interest in using ML techniques. Then, it slightly dropped from 23 studies to 22 in 2021.

Out of the ninety-four selected studies, sixty-three were journal articles, and thirty-one were conference papers. The most targeted publication source is Scopus for both journal articles and conference papers. However, IEEE and ScienceDirect have few related studies compared to Scopus; ten were selected from each. The large difference between the publication sources may be due to the duplication of studies when selecting the primary studies. Since Scopus was selected first, all the duplicated studies were removed. Thus, this may explain the high numbers of selected studies in Scopus. Figure 4 illustrates the publication trend and the distribution of primary studies among the selected publication sources.

B. RQ2. WHAT ARE THE MOST USED ML APPROACHES TO IMPUTE MISSING VALUES?

The answers to this question are obtained from the data item category ML technique, which includes the following subcategories: type of method used, single or hybrid, ML type, and selection method, as shown in Table 5.

The first subcategory from the data item extraction form determines what type of ML technique imputes the missing values. Every technique has its own characteristics and behavior based on the dataset used and the missing data mechanism. The selection of ML approaches is conducted based on the type of learning algorithm used, which includes supervised, unsupervised, and semisupervised or reinforcement learning. During the selection phase of primary studies, we noticed that some of the studies that included ML imputation as part of their proposal focused only on improving the prediction performance of some cases. However, these studies usually do not provide sufficient information about the data imputation process. Hence, these studies were omitted from our list by applying the fifth exclusion criteria in this study.

Table 7 demonstrates the selective primary study distribution over multiple ML approaches to impute missing values. The results show that hybrid ML methods are the most common among other approaches. This may be due to uncertainty when dealing with different data types in different missingness mechanisms. In addition, some ML algorithms require high computational resources while addressing missing value problems. Therefore, using hybrid methods can overcome this issue and increase the training time while simultaneously improving the performance. Hybrid ML models integrate two or more ML methods with each other or with other techniques to achieve higher performance. In this regard, some of the hybrid models used one method for prediction and the other one to optimize the prediction method to reach a new level of accuracy [105].

Hybrid models have recently become popular due to their high capability and potential, which explains the high focus on them in recent studies. A hybrid method can also combine supervised and unsupervised ML approaches to maximize data imputation performance [75]. The study proposed by Nikfalazar *et al.* [78] integrated a decision tree and fuzzy clustering to form a hybrid iterative model. The proposed model outperformed existing imputation methods in terms of computational speed while dealing with different data types simultaneously.

The second most commonly used ML method is deep learning. Based on artificial neural networks, deep learning has become popular in recent years due to the rapid advance in technology. Deep learning depends heavily on graphical processing units (GPUs), which accelerate the computational process in deep learning networks. Deep learning methods have been used in several domains, mainly for prediction, and have proven to perform well in imputing missing values, especially in high-dimensional datasets [55], [106].

The other ML approaches used in the selected studies include Clustering, Neural Network, XGBoost, K-Nearest Neighbor, and Ensemble methods. Most clustering studies used fuzzy theories to impute missing values because they can handle uncertainty, imprecision, and unevenness in several applications [35]. Liu *et al.* [32] used a fuzzy membership function to impute missing values under MNAR settings with the help of K-nearest neighbor to speed up the process provided with an iterative step to utilize historical data. The proposed study has proven its capability in handling missing values when there is a weak relationship between

TABLE 7. ML approaches used in imputing missing values.

Index	Name	Number of studies	References
1	Neural Networks	8	[22], [23], [24], [25], [26], [27], [28], [29]
2	Association Rule	1	[30]
3	Bayesian Networks	1	[31]
4	Clustering	9	[12], [32], [33], [34], [35], [36], [37], [38], [39]
5	Performance Comparison	11	[40], [41], [42], [43], [25], [27], [44], [31], [45], [46], [47], [48]
6	Decision Tree	2	[49], [50]
7	Deep Learning	15	[51], [52], [53], [33], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64]
8	Ensemble	5	[65], [66], [10], [67], [68]
9	Gaussian Process	1	[69]
10	Genetic Algorithm	1	[70]
11	Hierarchical Supervised Imputation Method	1	[71]
12	Hybrid	17	[72], [73], [74], [75], [20], [76], [19], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86]
13	Inductive Learning Algorithm	1	[21]
14	K-Nearest Neighbor	5	[87], [88], [20], [89], [90]
15	Missforest	2	[91], [92]
16	Principal Component Analysis	1	[93]
17	Reinforcement learning	1	[94]
18	Support Vector Machine	2	[95], [96]
19	Tensor-based	2	[97], [98]
20	Two-step ML imputation	1	[99]
21	eXtreme Gradient Boosting	5	[100], [101], [102], [103], [104]

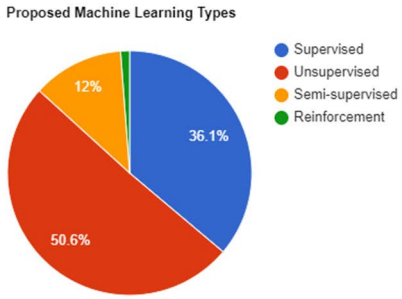

FIGURE 5. The proposed ML imputation methods types.

attributes. Nishanth *et al.*[22] proposed a probabilistic neural network imputation method preceded by mode to address multiple missing values from the categorical data type. Another study proposed by Madhu *et al.* [101] used extreme gradient boosting (XGBoost) to impute continuous and discrete missing data attributes in health care datasets from observable data. This study shows that XGBoost can effectively impute mixed-type missing values and achieve higher accuracy than conventional imputation methods.

Generally, ML approaches are categorized under four types: supervised, unsupervised, semi-supervised, and reinforcement learning. As shown in Figure 5, half of the studies followed the unsupervised learning type, which explores hidden patterns in data. Meanwhile, supervised learning studies came in second place, constituting approximately 36% of all primary studies.

In supervised learning, ML algorithms are trained to predict and classify data in a labeled dataset. As we can see, few semisupervised approaches were utilized for data imputation. This learning method stands between both supervised and unsupervised. Sometimes it combines clustering as a preprocessing step and classification using any of the supervised learning algorithms. Some of the proposed hybrid models use the semisupervised technique to achieve higher accuracy. Last but not least is reinforcement learning, which does not seem that popular in data imputation context, with one study only using it for optimization purposes.

Notably, combining multiple ML methods can markedly enhance the performance of imputing missing values. Table 8 illustrates a portion of the primary studies that compared different ML approaches to find the best-performing model that efficiently addresses the missing values problem. These results show that the K-nearest neighbor was the best performing method compared to other ML approaches. The study proposed by Huang *et al.* [25] shows that K-nearest neighbor produced the highest accuracy when combined with the genetic algorithm, for instance, selection. More studies show that combining K-NN with other ML methods, such as decision tree and random forest, led to better performance than other approaches [31], [40].

Moreover, from the rest of the comparison studies, we can see that random forest was the best performer three times, and the support vector machine came in third place, being the best performer twice. Another significant finding found in the study conducted by Nwulu [43] shows that multilayer perceptron (a feedforward artificial neural network) outperformed support vector machine, which achieved the highest accuracy in other studies. This indicates that neural networks and even deep networks can provide promising results regarding missing data imputation compared to existing ML approaches.

TABLE 8. Comparison studies of ML approaches.

Index	References	Number of Approaches	Outperforming Approach
1	[40]	5	Decision tree with K-NN
2	[41]	5	Random Forest
3	[42]	5	Support Vector Machine
4	[43]	3	Multi-Layer Perceptron
5	[25]	3	K-Nearest Neighbor
6	[27]	5	Support Vector Machine
7	[44]	20	Vector Autoregression
8	[45]	5	Variational Autoencoder
9	[46]	3	K-Nearest Neighbor
10	[47]	5	Random Forest
11	[48]	8	Inverse Distance Weighting
10	[31]	4	Random Forest with K-NN

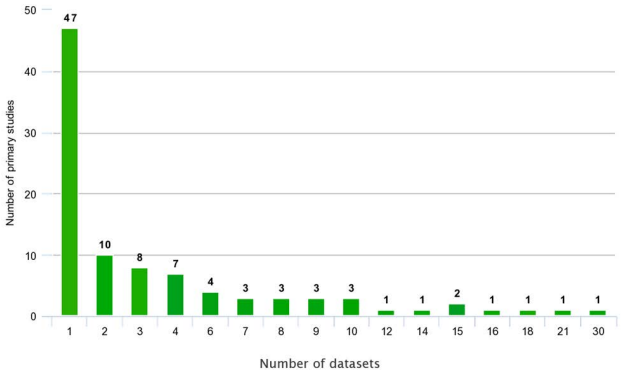

FIGURE 6. Distribution of primary studies by the number of datasets used for data imputation.

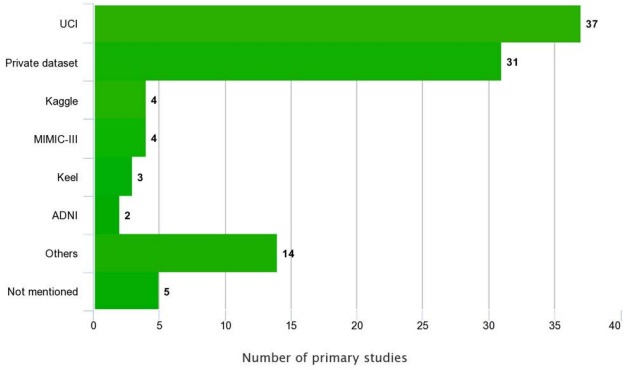

FIGURE 7. Distribution of primary studies by dataset sources.

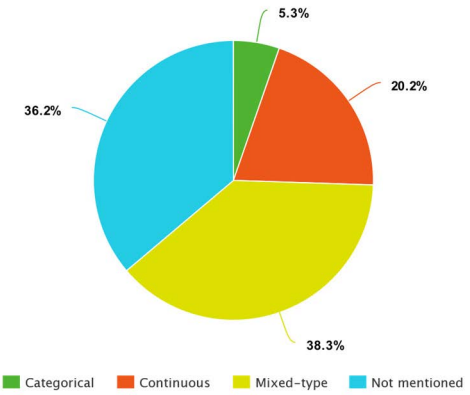

FIGURE 8. Percentage of the data types used for data imputation in primary studies.

C. RQ3. WHAT ARE THE CHARACTERISTICS OF DATASETS AND TYPES OF ANALYSIS USED IN THE PRIMARY STUDIES?

The answer to this research question is extracted from the Table 5 dataset category: number of datasets, dataset sources, data types, missing values ratio, missingness mechanism, and the type of analysis.

The first subcategory in the dataset identifies the number of datasets used in each primary study. The reviewed studies used various datasets in data imputation, ranging from 1 to 30 collected datasets. Figure 6 illustrates the number of datasets used in primary studies. Additionally, it was observed that most of the studies used only one dataset to deal with missing values (47 studies), whether they were synthetically generated or real. Furthermore, fewer studies tend to use more than one dataset since most of them deal with real-world datasets.

The second item in the dataset category is the source of the collected datasets. Figure 7 demonstrates the number of primary studies by dataset source. In total, 37 out of 94 primary studies collected the dataset from the University of California at Irvine (UCI) ML repository and utilized it for conducting data imputation experiments. A total of 31 primary studies used a private dataset that is not publicly available to evaluate their imputation method on a real-world scenario. For example, Chivers *et al.* in [99] proposed a two-step ML imputation method for a precipitation time-series dataset collected from 37 weather stations in the UK. Another study by Tavazzi *et al.* in [90] proposed an imputation algorithm to handle missing values in a clinical, epidemiological register of patients from two Italian regions. The other public dataset repositories used in primary studies are Kaggle, MIMIC-III, Keel, ADNI, etc. However, 5 out of 94 primary studies did not provide any information about their collected dataset.

After investigating missing value data types that are imputed in primary studies, it was found that 20.2% of the studies applied ML imputation methods on continuous data types. On the other hand, only 5.3% of studies focused on categorical data type only. Most of the studies (38.3%) dealt with numerical and categorical data types. However, 36.2% of the proposed methods did not provide any information about the data types imputed using their proposed method. Figure 8 shows the proportion of data types imputed in primary studies.

Moreover, the missing values ratio addressed in primary studies was also investigated. There are two types of missing values: real missingness that was missed naturally and artificial missingness generated synthetically for evaluation purposes. Among 94 studies, only one study used both real and artificial missingness types in their experiments. A total of 34 studies generated missing values in different ratios,

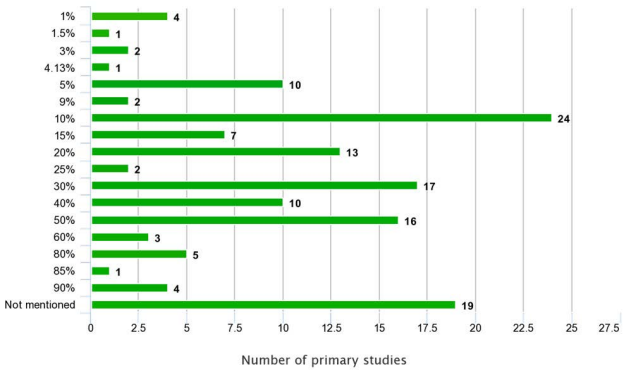

FIGURE 9. Percentage of missing values addressed in primary studies.

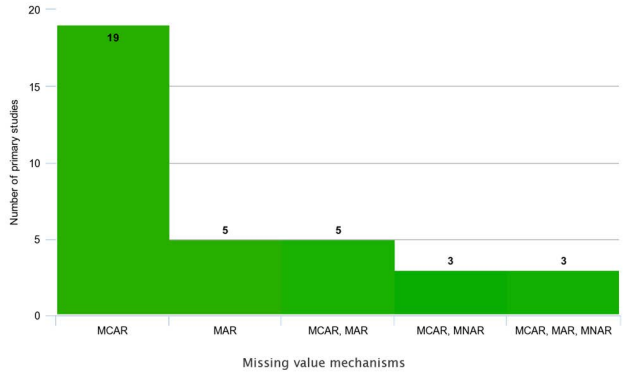

FIGURE 10. Distribution of primary studies by the missing value mechanism handled.

including 10% and 30%, which were the most commonly used among the other ratios. The highest missingness ratio used in the experiments was 90%, and the lowest was 1%. However, only 6 (6.4%) studies used datasets with real missingness. Furthermore, 19 out of 94 studies did not state any information about the missingness percentage. Figure 9 depicts the missing value ratio used in the primary study experiments.

The missing values mechanism is another item that was investigated in the primary studies. Most of the studies have provided a brief description of the different types of missingness mechanisms, including MCAR, MAR, and MNAR. We noticed that 63% of the primary studies did not indicate the type of missingness mechanism to be handled, whether it was artificially imposed or tested statistically on the dataset [107]. In total, 32% of the proposed imputation methods in primary studies used MCAR settings to impute missing values, and only five studies used the MAR mechanism alone. However, the procedure for handling the MNAR mechanism is complicated, and as shown in Figure 10, fewer imputation approaches can address it. Overall, the most commonly used mechanism is MCAR due to the ability to generate it artificially and test it statistically. Figure 10 represents the missing value mechanisms mentioned in primary studies.

It was observed that different types of analysis are used to deal with the dataset. A total of 12 out of 94 studies adapted their proposed ML imputation method to handle time-series data, including a collection of observations for a single entity at different time intervals. Missing values in time series data can occur due to multiple reasons, including data transmission error, measurement faults, and incorrect installation [61]. Oehmcke *et al.* in [72] developed an imputation algorithm that utilizes the correlation between features to obtain distance weights and imputes missing values consecutively. The authors claim that their approach shows accurate results on datasets with a high correlation between their features. Deep learning techniques were also adopted to handle time-series data. Liu *et al.* in [51] developed a deep learning method that captures the correlations between deep layers and the initial layers. Körner *et al.* [100] compared different ML imputation methods to impute meteorological time series data. The results show that XGBoost outperformed other imputation methods without considering the correlations between dataset features. Zhang *et al.* [92] showed that missForest could impute missing values in time series data with high accuracy regardless of the gap rate, unlike traditional methods, which lose their accuracy as the gap size of data increases. Another comparative study performed by Velasco-Gallego *et al.* in [44] between ML-based imputation methods and time series models to assess their ability to handle missing values in real-time. The authors applied a time series cross-validation type of analysis to form the training set based on the test set's prior measurements, including missing values. Phan in [77] proposed an ML method to impute missing values in univariate time series data using backward and forward forecasting based on historical values. The proposed method explores all the available values for the selected variable to estimate the missing values.

Most of the primary studies used a cross-sectional type of analysis for datasets collected at a single point in time. Unlike cross-sectional data, time-series data include autocorrelation, representing the degree of similarity between a given time series and its lagged version. Thus, handling missingness in time series data should be done carefully [34].

D. RQ4. WHAT ARE THE MEASUREMENT FACTORS USED TO EVALUATE THE ML IMPUTATION METHOD?

The answer to this research question is derived from Table 5 Performance Evaluation category: Performance metrics and evaluation methods.

We have investigated the performance metrics used to evaluate ML techniques. Table 9 shows the performance measures that are used more than once in primary studies to minimize the list as much as we can and focus on the important and trending factors.

From Table 9, the root mean square error (RMSE), also known as the root mean square deviation (RMSD), is the most commonly used factor to measure the performance of ML methods. It is calculated by subtracting the values predicted by the proposed imputation model from the observed values

in the dataset. The difference between these values is called residuals, and RMSE combines them into a single measurement value to estimate their prediction potency [65]. The second most widely used evaluation metric is accuracy, which is usually called classification accuracy, referring to classification approaches used in studies. The third most used measurement factor is the mean absolute error (MAE) or mean absolute deviation (MAD), which signifies the negative and positive deviation between the estimated and observed variables [41]. As noted in the literature, the RMSE value is always equal to or greater than the MAE. Both of these measurement metrics can range between zero and infinity. The larger the difference between them, the higher the variance of errors in the dataset. Next is the execution time, which varies among studies based on the size of the dataset, missing values, and computational resources used to conduct the experiment [74]. Huang *et al.* [87] indicated that the complication in their proposed method causes a significant increase in execution time for the algorithm. However, the method provided better prediction accuracy than existing methods. To reduce the time consumed by the model, Gupta *et al.* [112] highlighted that several factors related to the dataset could influence the execution time either positively or negatively.

Moreover, several performance metrics, including accuracy, recall, precision, specificity, and F-score, used the confusion matrix to calculate the performance of a classification model based on actual and predicted values. All of these measurements were used in five or more primary studies. Table 10 shows the confusion matrix with four different results: true positive (TP), true negative (TN), false-positive (FP), and false-negative (FN).

Referring to the confusion matrix, it is observed that both Recall and Sensitivity have the same calculation (TP/TP+FN). According to Table 9, recall was mentioned in six different studies, and sensitivity was mentioned in eight studies. Hence, the total number of studies using the same measurement was 14, which is equal to the number of studies that used mean absolute error as a performance measurement.

Moreover, the evaluation methods used to assess the proposed methods were scrutinized. This data item was extracted from the selected primary studies and was used to determine the level of adequacy for primary studies in evaluating their proposed methods. Table 11 summarizes the assessment methods used to validate ML techniques in imputing missing values. The results found two different evaluation methods: the experimental approach and the case study. The experimental approach usually addresses one or more cases to assess the capability of the proposed imputation methods in handling missing values across different settings. On the other hand, case studies only validate the proposed approach using a single case. It was found that most of the studies (81%) followed the experimental approach to validate their proposed ML method.

More investigation was conducted with regard to the domain of the proposed ML imputation methods.

TABLE 9. Measurement factors used in primary studies.

Index	Performance Measure	Number of times used	References
1	Area Under the Curve (AUC)	8	[71], [108], [93], [109], [55], [69], [90], [64]
2	Accuracy	36	[40], [42], [43], [73], [87], [88], [110], [111], [24], [25], [30], [52], [32], [75], [20], [95], [27], [13], [19], [21], [89], [93], [98], [101], [109], [31], [49], [56], [57], [59], [99], [104], [50], [96], [62], [38], [29]
3	Coefficient of determination (R^2)	6	[41], [99], [82], [46], [84], [86]
4	Execution time	10	[40], [74], [87], [26], [21], [93], [98], [44], [49], [56]
5	F-score	7	[65], [41], [23], [100], [99], [96], [64]
6	Kappa statistics	3	[95], [109]
7	Mean Absolute Error (MAE):	16	[12], [41], [23], [20], [100], [108], [112], [58], [35], [69], [70], [77], [78], [37], [82], [47], [92]
8	Mean Square Error (MSE)	3	[43], [49], [60]
9	Mean-relative error (MRE)	2	[97], [58]
10	Normalized absolute error (nAE)	2	[79], [90]
11	Normalized root mean square error (nRMSE)	10	[32], [19], [91], [98], [102], [55], [68], [63], [92], [64]
12	Precision	11	[71], [73], [111], [20], [27], [55], [90], [99], [104], [64], [38]
13	proportion of false classification (PFC)	5	[91], [55], [90], [63], [64]
14	Recall	9	[111], [95], [55], [90], [99], [104], [50], [64], [38]
15	ROC (receiver operating characteristic curve)	9	[73], [111], [20], [108], [27], [98], [69], [90], [50]
16	Root Mean Square Error (RMSE)	38	[65], [41], [51], [74], [87], [23], [75], [20], [100], [112], [26], [54], [34], [93], [97], [101], [44], [58], [35], [36], [59], [70], [77], [78], [80], [37], [99], [45], [82], [61], [83], [46], [62], [47], [84], [28], [29], [39]
17	Sensitivity	8	[40], [71], [52], [19], [93], [98], [109], [69]
18	Specificity	6	[52], [19], [93], [98], [109], [69]
19	Variance of error (VARE)	2	[41], [101]

Figure 11 illustrates the allocation of primary studies over their domain. The results show that approximately 64% of the primary studies used a specific domain to conduct

TABLE 10. Confusion matrix.

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive	False-Positive
	Negative	False Negative	True Negative

TABLE 11. Evaluation methods to validate the proposed ML techniques.

Index	Evaluation Method	Number of studies	References
1	Experimental Study	76	[22], [12], [71], [65], [66], [10], [72], [94], [42], [43], [67], [73], [87], [88], [110], [24], [25], [30], [52], [53], [32], [75], [20], [95], [108], [112], [26], [27], [33], [13], [76], [19], [21], [89], [91], [97], [98], [101], [102], [109], [31], [49], [55], [56], [57], [35], [36], [69], [59], [70], [77], [78], [79], [80], [81], [37], [90], [103], [104], [50], [45], [60], [83], [68], [46], [96], [62], [63], [47], [84], [64], [28], [85], [38], [29], [86]
2	Case Study	18	[40], [41], [51], [74], [111], [23], [100], [54], [34], [93], [44], [58], [99], [82], [61], [92], [48], [39]

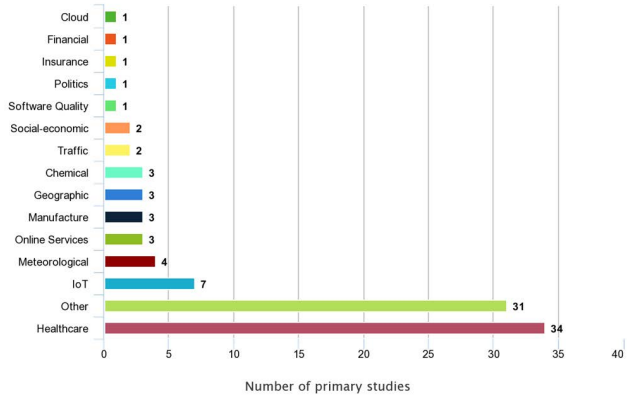

FIGURE 11. Domain of the existing ML imputation methods.

data imputation. However, the remaining primary studies dealt with multiple domains and were not dependent on any specific domain.

Most primary studies used health care as their domain to evaluate the proposed ML imputation method. (34 out of 94). This domain includes the medical, clinical, and bioinformatics fields. Several ML approaches have been utilized in the health care field, such as deep learning, XGBoost, K-NN, ensemble, and hybrid methods. These algorithms were used to impute missing values, while XGBoost was used to bypass them and produce high-performance results. The second rank in the most used domain list is IoT, in which deep learning techniques were used 3 times out of the 7 primary studies, and three of them followed the case study evaluation method. One-third of the listed domains were used only once by the primary studies, while four domains, including online services, geographic, manufacturing, and chemical domains, were used three times.

Meanwhile, four studies used the case study evaluation approach to propose their ML imputation methods in the meteorological domain. The former was used two times in case studies for socio-economic and traffic domains, while the latter was used only in experimental studies. Moreover, a large portion of primary studies (33%) used multiple datasets from different domains to evaluate the performance of their proposed ML imputation method.

Different ML imputation methods were applied in different settings depending on their characteristics. Usually, authors may not possess sufficient knowledge in the domain they are working on, which applies to the data domain they are also using. As has been observed, most researchers tend to use deep learning approaches because of their ability to deal with high-dimensional data despite a lack of knowledge in the chosen domain. Another important factor to be considered for selecting an ML approach is its characteristics. For instance, the K-nearest neighbor is widely used because of its simplicity and high performance. However, K-NN performs poorly in large datasets and high-dimensional data contexts [47]. Additionally, many researchers stumble while linking different features in a dataset. Hence, knowing the domain needs to be considered to improve research results.

E. RQ5. WHAT ARE THE LIMITATIONS AND STRENGTH POINTS IN APPLYING ML METHODS TO IMPUTE MISSING VALUES?

To answer this research question, we extracted the data item limitations and strengths according to Table 5. It was found that less than half of the studies (40%) clearly stated the strength of the proposed ML imputation approach. Meanwhile, 42% of the studies described the limitation of the proposed approach.

The strength of the proposed ML approaches was occasionally reported in the studies. In most studies, researchers have indicated that ML imputation methods can deal with high missingness ratios while maintaining a small RMSE, unlike traditional imputation methods such as deletion, mean substitution, or Hot-Deck [31], [33], [51], [72], [75], [110]. Other studies affirmed that using ML in imputing missing values has gradually improved accuracy and provided more accurate results regardless of dataset size [19], [31], [67], [73], [97].

ML approaches also proved the capability of fast learning by taking advantage of the current advanced computational resources. Nagarajan and Dhinesh [76] declared that the multiple imputation method suffers from high computational consumption, which leads to the use of ML to bypass this limitation. Currently, ML algorithms can impute missing values in large datasets within a couple of minutes compared with other imputation methods, such as MICE, which can take several hours to impute the same dataset [10], [79]. Moreover,

several researchers claimed that their proposed ML methods could be generalized and applied to other domains where data imputation is needed [10], [87], [90]. In addition, other studies concluded that ML methods could handle uncertainty and vagueness by extracting more information from datasets that contain noisy data in a large-scale context [35], [59], [99].

Among all ML algorithms, deep learning imputation techniques have shown notable results by modeling complex relationships between variables in a dataset [55]. The work proposed by Cheng *et al.* [56] shows that deep learning methods can impute the whole sample instead of focusing on only one group at a time. It is interesting to note that all the studies highlighting the strength of deep learning models were conducted in the health care domain.

Unlike conventional imputation methods, ML approaches incorporate a unique learning strategy that considers links and relationships among sparse variables to improve the data imputation performance [55]. The imputation method proposed by Hu and Du [69] demonstrates the ability to learn from the correlations between features to estimate and impute the missing values. In contrast, a study by Liu *et al.* [32] indicated that the proposed approach performed better when there were weak correlations between features. Other studies argued that conducting preprocessing, including instance selection first, improves the data imputation performance of all different datatypes of the selected dataset [25], [100].

Although several studies reported strength points, there is an almost similar number of studies that stated limitations of the proposed ML imputation methods. One of the reported limitations is the slow training speed of some ML approaches, such as generative adversarial networks (GANs) and stochastic gradient boosting machines (SGBMs) [33], [91]. Both of these methods are more computationally intensive than multiple imputation and missForest. Other studies did not consider all missing data mechanisms and only tested the MCAR scenario [62], [93]. A study by Vilardell *et al.* [109] acknowledged that the data quality of the dataset could directly affect the performance of the ML imputation method.

Additionally, limitations in datasets, including size, dimensionality, and diversity, can greatly influence the results. Peralta *et al.* [60] conducted a study on a low number of complete data points, and their proposed imputation method was not tested with the categorical data type. Li *et al.* [61] mentioned that their method should be evaluated on other data types to verify its effectiveness for generalization. Xu *et al.* [64] indicated that the limitation in dataset size and diversity could affect the imputation performance. In addition, missing ratios play a great role and require more attention because they can also affect the proposed method's performance [55]. Marshall *et al.* [113] suggested that a missing ratio over 50% is not acceptable in the health care domain. Hence, clinical researchers should be aware not to exceed this threshold while simulating missing values. Furthermore, outliers can also affect the imputation performance by causing a large RMSE if not handled correctly by deleting or replacing them [47].

It is important to note that deep learning imputation methods learn better when the dataset size is more than 1,000 instances. However, if more iterations are conducted, it can reduce predictive accuracy power [56]. Methods using self-organizing maps also suffer from overfitting problems when the number of nodes increases, leading to lower accuracy [73]. Kim and Chung [57] affirmed this limitation in their study and stated that even computational speed decreases as the relationships among variables are learned. If there are weak dependencies among features, imputed data may lack precision [79]. Generally, ML techniques are highly sensitive to uncertainty in data, and small variations in input data may lead to remarkable changes in output data [69].

Furthermore, ML algorithms do not perform equally on comparable datasets, but they may vary depending on the features available for analysis [99]. Last, the generalization of ML imputation techniques is not applicable in all scenarios. If the number of variables is larger than the tested dataset, feature selection should be considered to drop highly correlated attributes [103].

V. DISCUSSION

This section discusses the findings of the systematic literature review. In addition, recommendations and future work directions of ML imputation methods are provided.

A. FINDINGS

The main findings of this review include the following:

1) THE CONTRIBUTION OF ML TRENDS IN IMPUTING MISSING VALUES

While answering the first research question and analyzing the extracted data items from Table 5, it was found that implementing ML approaches in imputing missing values has increased in the last 6 years (2016-2021). This is considered one of the contributing factors in the data quality domain. ML is usually used to predict the missing values in different dataset settings, predict the labeled class values in supervised learning, detect hidden patterns, and group them in unsupervised learning. Furthermore, high-dimensional datasets and the sparsity in large data with different datatypes and missingness mechanisms also contribute to ML trends. These challenges have triggered researchers to improve algorithms to find the best solution for missing values. Subsequently, ML methods have provided remarkable results in dealing with this issue.

These studies have also shown that half of the primary studies used unsupervised ML techniques. However, the most commonly used ML approach is the hybrid model, which usually has the advantage of both supervised and unsupervised learning techniques in handling prediction and optimization at the same time. Furthermore, ML provides better decision-making by learning from observed values and estimating missing values accurately. Additionally, ML has

proven to have the ability to deal with uncertainty by detecting and extracting noisy data dynamically in high-dimensional datasets. Tavazzi *et al.* [90] indicated that ML techniques could extract information from a large and complex amount of data in several domains.

2) THE LACK OF FOCUSED DOMAINS

One of the key findings in this study is that there is a high bias in focusing on the health care domain to address the missing values problem using ML imputation methods. However, most listed domains had only one study done, and the highest number of studies in domains other than health care is seven. In addition to health care, it was observed that a considerable number of studies dealt with multiple datasets from different domains to evaluate the performance of the proposed ML method. From these results, it is clear that most studies neglected their domain knowledge, and ML techniques were chosen simply based on their previous performance. Thus, understanding the selected domain characteristics and data requirements is important to improve data imputation performance.

3) HIGH DIVERSITY IN PERFORMANCE MEASUREMENT

Another finding in this research is the high diversity in performance measurement factors summarized in Table 9. We listed 19 different metrics used in primary studies that were used more than twice to avoid redundancy. The most commonly used factor was RMSE, but it was used in only 38 studies, constituting less than half of the primary studies. On the other hand, the computational time consumed for training the ML model was used in only ten studies. These two metrics show the degree of efficiency the model can reach since RMSE is used in most studies, and it should not be ignored.

Moreover, one of the main issues in existing methods, such as multiple imputations, is the high consumption of computational resources, which leads to a long execution time for imputing missing values. However, few studies have considered this an important performance factor. In other words, the high diversity in performance metrics creates more challenges in evaluating the proposed ML approaches with other related studies.

B. FUTURE RESEARCH DIRECTIONS

Based on the findings of this study, research directions are identified to improve the implementation of ML techniques to impute missing values.

1) TAXONOMY OF ML IN IMPUTING MISSING VALUES

ML approaches can replace missing values in datasets and significantly improve data quality and decision-making. However, each ML method has its advantages and disadvantages while dealing with specific domains in addition to datatype constraints decreed by these methods. Therefore, each ML algorithm is selected based on its characteristics and the ability to deal with collected datasets. For instance, deep learning has become the most commonly used approach because of its ability to handle complex relationships between data and impute missing values in large datasets. Generally, deep learning approaches impute missing values with high accuracy and short execution time. In this respect, directions are needed for adopting the best ML method to improve data imputation. Hence, our first future research direction is to propose a taxonomy that provides the procedure for selecting the most appropriate ML approach based on its features and domain type. Figure 12 illustrates the proposed taxonomy for selecting the best ML imputation method.

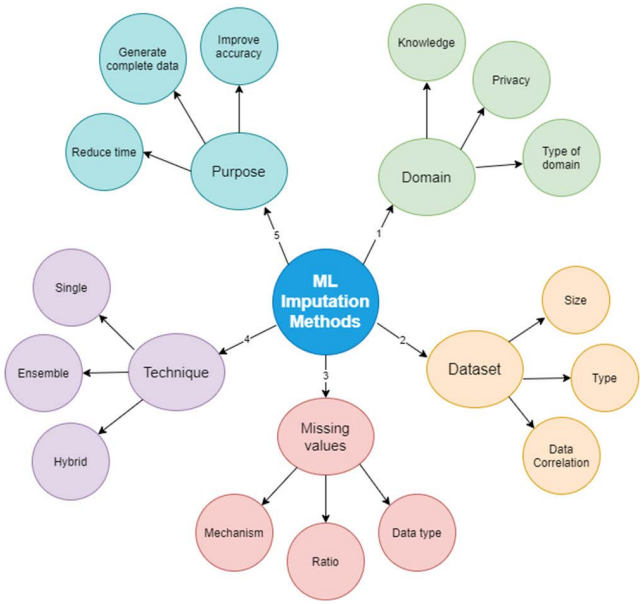

FIGURE 12. The proposed taxonomy for selecting the best ML imputation method.

Understanding domain requirements and settings are crucial to choosing the most suitable ML imputation approach because of the variety in its domain characteristics. One of its main characteristics is in the domain of knowledge, such as the advantages and disadvantages, the amount of data, and the comprehensive explanation needed. Next, is the privacy concern, which is more often found in the health care field, in protecting patients' sensitive data. Thus, domain privacy can limit the amount of data that can affect the choice of the ML approach. Moreover, the type of domain defines what kind of methods can address the missing values problem by clearly understanding the attributes and features.

Another important factor is the dataset, which is specified by its size, type, and level of correlation between variables. For instance, deep learning techniques work better with large datasets, while support vector machines and K-NN achieve high performance in small datasets [47]. Different datasets have different data types, such as time-series datasets, which work better with recurrent neural networks [108]. The level of correlation among variables can also affect the selection process of the ML method. The work proposed by [32] indicates that the proposed Fuzzy C-Means method performs well with weak correlations, while deep learning

and Gaussian process approaches provide better accuracy with highly correlated attributes.

The main focus of implementing ML is to impute missing values, and these missing values can differ from one dataset to another. Three main factors greatly affect the performance of the proposed ML approach, including missing value data types, missingness ratio, and the type of mechanism. Data types can range between numerical and categorical, which can be of binary or nominal value. However, some ML methods have constraints and cannot deal with all data types. Usually, ML methods can deal with high missingness ratios, especially hybrid techniques. The last factor regarding missing values is the type of mechanism. It is important to differentiate between MCAR, MAR, and MNAR when dealing with missing values. Although most ML approaches can address the MCAR and MAR settings, the latter mechanism is quite challenging.

ML techniques are categorized into three main types: single, ensemble, and hybrid. The latter is the most commonly used method in primary studies because it can deal with large datasets without consuming much time. Alternatively, computational power has improved dramatically in recent years. Currently, it is possible to add more ensembles to the proposed model to achieve higher performance [73]. The last technique is the single method, which can be supervised, unsupervised, semisupervised, and reinforce learning. The deep learning method is the most commonly used single approach in primary studies due to its high performance and scalability.

The last factor in our proposed taxonomy is the purpose of the data imputation process. Some of the studies aimed to improve accuracy, and as a result, methods such as XGBoost can fit in such scenarios due to their ability to ignore missing values and achieve high accuracy. However, other studies aimed to improve data completeness by estimating the missing values and generating a complete dataset for future analysis. In this case, other ML approaches should be considered, such as missForest. Reducing the training time can also be considered, especially when dealing with large datasets. Some ML approaches, such as deep learning and hybrid models, can significantly reduce the execution time. Therefore, future research should utilize the latest technology to speed up the training time.

2) INVESTIGATING MISSING NOT AT RANDOM PATTERN

The survey results found that the studies usually followed two main scenarios in imputing missing values. The first scenario is to generate synthetic missing values under a specific mechanism and control the missingness ratio. Several proposed ML imputation methods were tested under these settings to test their robustness in dealing with high missing values. The second scenario addresses missing values in a collected dataset without any manipulation. This scenario is usually followed while dealing with real-time data collected from sensors.

Most primary studies applied ML to impute missing values under the MCAR case since it is the easiest to address, while some studies dealt with MAR as well. However, most studies did not consider the MNAR mechanism, which includes missing values related to observable or nonobservable variables. According to Karanikola and Kotsiantis [13], missing values under the MNAR type are difficult to handle and have been avoided in several studies. Thus, it is strongly recommended that a robust ML method be developed to efficiently impute such values, thus making this solution attractive to other imputation problems and domains.

3) DEEP LEARNING ALGORITHMS

Concerning the limitations of the selected studies, ML approaches depended on computational resources more than other imputation methods. This is due to the training process that is needed in the ML algorithm to predict missing values. The consumption of computational resources could delay imputing missing values, especially when handling complex relationships among variables.

To address this issue, it is recommended that deep learning techniques be utilized. These techniques can deal with high-dimensional data and impute missing values in the whole sample instead of categorizing it into groups. Deep learning is the current trend in data imputation, as it is the most commonly used approach among other ML techniques in the last five years.

There are several deep learning approaches, each of which addresses different data types. For instance, recurrent neural networks are usually utilized to address missing values in longitudinal or time-series datasets [108]. Moreover, convolutional neural networks are another common technique that has achieved remarkable performance in the image classification field [57]. In case there is high uncertainty in dataset features, a generative adversarial network (GAN) is utilized, which also can deal with a high missingness ratio [110].

Deep learning models consist of multiple layers, including an input layer, hidden layers, and output layer. Every layer contains several nodes depending on the number of features in the input layer. These nodes are mapped nonlinearly from the input layer up to the output layer. Deep learning models can learn faster than other ML and traditional imputation methods. Therefore, deep learning can estimate missing values efficiently even if the data have complex relationships. In other words, deep learning has replaced multiple imputations by taking advantage of the new high-end computational resources.

Most of the primary studies used deep learning in the health care domain. The work proposed by Xu *et al.* [55] applied deep learning to address missing values in electronic health records. This study shows that deep learning methods consider the characteristics and relationship of patient data, including missing values and time-based patterns, which tend to be overlooked in other imputation techniques.

VI. CONCLUSION

The main objective of this systematic literature review was to evaluate the novel ML approaches in handling missing values. This study provides a comprehensive systematic literature review on ML imputation methods as proposed by 94 papers selected from the literature. In the reviewed literature, missing values were estimated and replaced using several ML techniques. The extracted data items were analyzed and highlighted to provide insights into current trends in ML imputation methods.

The existing studies show that the application of ML in imputing missing values has increased significantly over the last six years. Hybrid models and deep learning are the most commonly used methods in these studies, and half of the primary studies followed the unsupervised learning approach. Various factors were used to evaluate the performance of the proposed method. The most common factors used in the studies were RMSE and accuracy. Furthermore, the health care domain is found to have attracted more studies compared to other domains. The studies show that ML approaches can deal with high missingness rates while maintaining a small error regardless of the dataset size. This evolution in technology has positively impacted the performance of ML and dramatically increased its performance. Additionally, ML imputation methods have proven to be able to handle uncertainty and noisy data while considering links and relationships between data.

The limitations in the primary studies indicate that there is more work to be done to improve ML imputation methods in the future. Very few studies have used a systematic approach in selecting an ML approach for data imputation. Thus, we have proposed several recommendations that can be useful in improving data completeness by handling missing values in the most efficient way possible using state-of-the-art ML approaches. Other domains that can benefit from such approaches and need further exploration include computer-assisted language learning and intelligent virtual teaching assistants, particularly for online classrooms, which have become the new normal in educational contexts and teacher-student interactions.

ACKNOWLEDGMENT

All opinions, findings, conclusions, and recommendations in this article are those of the authors and do not necessarily reflect the views of the funding agencies. The authors would like to thank the anonymous reviewers for their comments.

REFERENCES

[1] R. Deb and A. W.-C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data," *Inf. Sci.*, vol. 339, pp. 274–289, 2016, doi: 10.1016/j.ins.2016.01.018.

[2] C.-F. Tsai and F.-Y. Chang, "Combining instance selection for better missing value imputation," *J. Syst. Softw.*, vol. 122, pp. 63–71, Dec. 2016, doi: 10.1016/j.jss.2016.08.093.

[3] K. Dhindsa, M. Bhandari, and R. R. Sonnadara, "What's holding up the big data revolution in healthcare?" *BMJ*, vol. 363, pp. 1–2, Dec. 2018, doi: 10.1136/bmj.k5357.

[4] M. Janssen, H. van der Voort, and A. Wahyudi, "Factors influencing big data decision-making quality," *J. Bus. Res.*, vol. 70, pp. 338–345, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.007.

[5] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.

[6] B. L. Ford, "An overview of hot-deck procedures," in *Incomplete Data Sample Surveys*, vol. 2. 1983.

[7] T. Thomas and E. Rajabi, "A systematic review of machine learning-based missing value imputation techniques," *Data Technol. Appl.*, vol. 55, no. 4, pp. 558–585, Aug. 2021, doi: 10.1108/DTA-12-2020-0298.

[8] T. D. Pigott, "A review of methods for missing data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, Dec. 2001, doi: 10.1076/edre.7.4.353.8937.

[9] J. Barnard, N. Schenker, and D. B. Rubin, *Multiple Imputation*, vol. 16, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2015.

[10] D. Sovilj, E. Eirola, Y. Miche, K. M. Björk, R. Nian, and A. Akusok, "Extreme learning machine for missing data using multiple imputations," *Neurocomputing*, vol. 174, pp. 220–231, Jan. 2016, doi: 10.1016/j.neucom.2015.03.108.

[11] A. Sadhu, R. Soni, and M. Mishra, "Pattern-based comparative analysis of techniques for missing value imputation," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 513–518, doi: 10.1109/ICCCA49541.2020.9250825.

[12] R. Razavi-Far and M. Saif, "Imputation of missing data using fuzzy neighborhood density-based clustering," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2016, pp. 1834–1841, doi: 10.1109/FUZZ-IEEE2016.7737913.

[13] A. Karanikola and S. Kotsiantis, "A hybrid method for missing value imputation," in *Proc. 23rd Pan-Hellenic Conf. Informat.*, Nov. 2019, doi: 10.1145/3368640.3368653.

[14] T. Köse, S. Özgür, E. Coşgun, A. Keskinoğlu, and P. Keskinoğlu, "Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study," *BioMed Res. Int.*, vol. 2020, pp. 1–15, Jul. 2020, doi: 10.1155/2020/1895076.

[15] B.-J.-M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, and K. M. Waters, "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics," *J. Proteome Res.*, vol. 14, no. 5, pp. 1993–2001, May 2015, doi: 10.1021/pr501138h.

[16] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Tech. Rep., 2007, doi: 10.1145/1134285.1134500.

[17] L. Chen, M. A. Babar, and H. N. Zhang, "Towards an evidence-based understanding of electronic data sources," in *Proc. 14th Int. Conf. Eval. Assessment Softw. Eng.* 2010, pp. 1–4, doi: 10.14236/ewic/ease2010.17.

[18] L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong, and D. Shao, "Quality assessment in systematic literature reviews: A software engineering perspective," *Inf. Softw. Technol.*, vol. 130, Feb. 2021, Art. no. 106397, doi: 10.1016/j.infsof.2020.106397.

[19] M. Vazifehdan, M. H. Moattar, and M. Jalali, "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 31, no. 2, pp. 175–184, Apr. 2019, doi: 10.1016/j.jksuci.2018.01.002.

[20] T. Nkonyana and B. Twala, *Impact of Poor Data Quality in Remotely Sensed Data*, vol. 668, Singapore: Springer, 2018.

[21] S. M. Abu-Soud, "A novel approach for dealing with missing values in machine learning datasets with discrete values," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Apr. 2019, pp. 1–5, doi: 10.1109/ICCISci.2019.8716430.

[22] K. J. Nishanth and V. Ravi, "Probabilistic neural network based categorical data imputation," *Neurocomputing*, vol. 218, pp. 17–25, Dec. 2016, doi: 10.1016/j.neucom.2016.08.044.

[23] A. Petrozziello, I. Jordanov, and C. Sommeregger, "Distributed neural networks for missing big data imputation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489488.

[24] K.-F. Jea, C.-W. Hsu, and L.-Y. Tang, "A missing data imputation method with distance function," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2018, pp. 450–455, doi: 10.1109/ICMLC.2018.8526985.

[25] M.-W. Huang, W.-C. Lin, and C.-F. Tsai, "Outlier removal in model-based missing value imputation for medical datasets," *J. Healthcare Eng.*, vol. 2018, Feb. 2018, Art. no. 1817479, doi: 10.1155/2018/1817479.

[26] Q. Ma, Y. Gu, W.-C. Lee, and G. Yu, "Order-sensitive imputation for clustered missing values," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 1, pp. 166–180, Jan. 2019, doi: 10.1109/TKDE.2018.2822662.

[27] E. T. Caparino, A. M. Sison, and R. P. Medina, "Application of the modified imputation method to missing data to increase classification performance," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Feb. 2019, pp. 134–139, doi: 10.1109/CCOMS.2019.8821632.

[28] N. Savarimuthu and S. Karesiddaiah, "An unsupervised neural network approach for imputation of missing values in univariate time series data," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 9, pp. 1–16, May 2021, doi: 10.1002/cpe.6156.

[29] K. Shobha and N. Savarimuthu, "Clustering based imputation algorithm using unsupervised neural network for enhancing the quality of healthcare data," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 1771–1781, Feb. 2021, doi: 10.1007/s12652-020-02250-1.

[30] S. A. Zahin, C. F. Ahmed, and T. Alam, "An effective method for classification with missing values," *Int. J. Speech Technol.*, vol. 48, no. 10, pp. 3209–3230, Oct. 2018, doi: 10.1007/s10489-018-1139-9.

[31] A. Idri, I. Kadi, I. Abnane, and J. L. Fernandez-Aleman, "Missing data techniques in classification for cardiovascular dysautonomias diagnosis," *Med. Biol. Eng. Comput.*, vol. 58, no. 11, pp. 2863–2878, Nov. 2020, doi: 10.1007/s11517-020-02266-x.

[32] S. Liu, H. Dai, and M. Gan, "Information-decomposition-model-based missing value estimation for not missing at random dataset," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 1, pp. 85–95, Jan. 2018, doi: 10.1007/s13042-015-0354-5.

[33] H. Wang, Z. Yuan, Y. Chen, B. Shen, and A. Wu, "An industrial missing values processing method based on generating model," *Comput. Netw.*, vol. 158, pp. 61–68, Jul. 2019, doi: 10.1016/j.comnet.2019.02.007.

[34] X. Sun, Z. Wang, and J. Hu, "ELM-PSO-FCM based missing values imputation for byproduct gas flow data analysis," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Mar. 2019, pp. 56–59, doi: 10.1109/ITNEC.2019.8729038.

[35] P. S. Raja, K. Sasirekha, and K. Thangavel, "A novel fuzzy rough clustering parameter-based missing value imputation," *Neural Comput. Appl.*, vol. 32, no. 14, pp. 10033–10050, Jul. 2020, doi: 10.1007/s00521-019-04535-9.

[36] C. Garcia, D. Leite, and I. Skrjanc, "Incremental missing-data imputation for evolving fuzzy granular prediction," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 10, pp. 2348–2362, Oct. 2020, doi: 10.1109/TFUZZ.2019.2935688.

[37] P. S. Raja and K. Thangavel, *Missing Value Imputation Using Unsupervised Machine Learning Techniques*, vol. 24, no. 6. Berlin, Germany: Springer, 2020.

[38] X. Hu, W. Pedrycz, K. Wu, and Y. Shen, "Information granule-based classifier: A development of granular imputation of missing data," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106737, doi: 10.1016/j.knosys.2020.106737.

[39] C. Velasco-Gallego and I. Lazakis, "A novel framework for imputing large gaps of missing values from time series sensor data of marine machinery systems," *Ships Offshore Struct.*, to be published., doi: 10.1080/17445302.2021.1943850.

[40] M. Askarian, G. Escudero, M. Graells, R. Zarghami, F. Jalali-Farahani, and N. Mostoufi, "Fault diagnosis of chemical processes with incomplete observations: A comparative study," *Comput. Chem. Eng.*, vol. 84, pp. 104–116, Jan. 2016, doi: 10.1016/j.compchemeng.2015.08.018.

[41] I. Gad and B. R. Manjunatha, "Performance evaluation of predictive models for missing data imputation in weather data," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 1327–1334, doi: 10.1109/ICACCI.2017.8126025.

[42] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, "Influence of data distribution in missing data imputation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Computer Science Artificial Intelligent Lecture Notes Bioinformatics)*, vol. 10259. 2017, pp. 285–294, doi: 10.1007/978-3-319-59758-4_33.

[43] N. I. Nwulu, "Evaluation of machine learning classification algorithms & missing data imputation techniques," in *Proc. Int. Artif. Intell. Data Process. Symp. (IDAP)*, Sep. 2017, pp. 1–5, doi: 10.1109/IDAP.2017.8090315.

[44] C. Velasco-Gallego and I. Lazakis, "Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study," *Ocean Eng.*, vol. 218, Dec. 2020, Art. no. 108261, doi: 10.1016/j.oceaneng.2020.108261.

[45] N. U. Okafor and D. T. Delaney, "Missing data imputation on IoT sensor networks: Implications for on-site sensor calibration," *IEEE Sensors J.*, vol. 21, no. 20, pp. 22833–22845, Oct. 2021, doi: 10.1109/JSEN.2021.3105442.

[46] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, J. Chen, M. A. Chyad, S. Garfan, and A. M. Aleesa, "Machine learning-based imputation soft computing approach for large missing scale and non-reference data imputation," *Chaos, Solitons Fractals*, vol. 151, Oct. 2021, Art. no. 111236, doi: 10.1016/j.chaos.2021.111236.

[47] Y. Fu, H. Liao, and L. Lv, "A comparative study of various methods of handling missing data in UNSODA," *Agriculture*, vol. 11, no. 8, p. 727, 2021, doi: 10.3390/agriculture11080727.

[48] R. Rodríguez, M. Pastorini, L. Etcheverry, and C. Chreties, "Water-quality data imputation with a high percentage of missing values: A machine learning approach," *Sustainability*, vol. 13, no. 11, p. 6318, 2021, doi: 10.3390/su13116318.

[49] C. Beaulac and J. S. Rosenthal, "BEST: A decision tree algorithm that handles missing values," *Comput. Statist.*, vol. 35, no. 3, pp. 1001–1026, Sep. 2020, doi: 10.1007/s00180-020-00987-z.

[50] D. Cenitta, R. V. Arjunan, and K. V. Prema, "Missing data imputation using machine learning algorithm for supervised learning," in *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, Jan. 2021, pp. 1–5, doi: 10.1109/ICCCI50826.2021.9402558.

[51] Z. Liu, W. Zhang, S. Lin, and T. Q. S. Quek, "Heterogeneous sensor data fusion by deep multimodal encoding," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 479–491, Apr. 2017, doi: 10.1109/JSTSP.2017.2679538.

[52] E. Jabason, M. O. Ahmad, and M. N. S. Swamy, "Missing structural and clinical features imputation for semi-supervised Alzheimer's disease classification using stacked sparse autoencoder," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2018, pp. 1–4, doi: 10.1109/BIOCAS.2018.8584844.

[53] C.-B. Lu and Y. Mei, "An imputation method for missing data based on an extreme learning machine auto-encoder," *IEEE Access*, vol. 6, pp. 52930–52935, 2018, doi: 10.1109/ACCESS.2018.2868729.

[54] G. Boquet, J. L. Vicario, A. Morell, and J. Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2882–2886.

[55] D. Xu, P. J. H. Hu, T. S. Huang, X. Fang, and C. C. Hsu, "A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management," *J. Biomed. Inform.*, vol. 111, Nov. 2020, Art. no. 103576, doi: 10.1016/j.jbi.2020.103576.

[56] C.-Y. Cheng, W.-L. Tseng, C.-F. Chang, C.-H. Chang, and S. S.-F. Gau, "A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder," *Frontiers Psychiatry*, vol. 11, pp. 1–13, Jul. 2020, doi: 10.3389/fpsyt.2020.00673.

[57] J.-C. Kim and K. Chung, "Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data," *IEEE Access*, vol. 8, pp. 104933–104943, 2020, doi: 10.1109/ACCESS.2020.2997255.

[58] F. Guo, W. Bai, and B. Huang, "Output-relevant variational autoencoder for just-in-time soft sensor modeling with missing data," *J. Process Control*, vol. 92, pp. 90–97, Aug. 2020, doi: 10.1016/j.jprocont.2020.05.012.

[59] M. Kachuee, K. Karkkainen, O. Goldstein, S. Darabi, and M. Sarrafzadeh, "Generative imputation and stochastic prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1278–1288, Mar. 2022, doi: 10.1109/TPAMI.2020.3022383.

[60] M. Peralta, P. Jannin, C. Haegelen, and J. S. H. Baxter, "Data imputation and compression for Parkinson's disease clinical questionnaires," *Artif. Intell. Med.*, vol. 114, Apr. 2021, Art. no. 102051, doi: 10.1016/j.artmed.2021.102051.

[61] J. Li, W. Ren, and M. Han, "Variational auto-encoders based on the shift correction for imputation of specific missing in multivariate time series," *Measurement*, vol. 186, Dec. 2021, Art. no. 110055, doi: 10.1016/j.measurement.2021.110055.

[62] Y. Wang, D. Li, X. Li, and M. Yang, "PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data," *Neural Netw.*, vol. 141, pp. 395–403, Sep. 2021, doi: 10.1016/j.neunet.2021.05.033.

[63] W. Dong, D. Y. T. Fong, J.-S. Yoon, E. Y. F. Wan, L. E. Bedford, E. H. M. Tang, and C. L. K. Lam, "Generative adversarial networks for imputing missing data for big data clinical research," *BMC Med. Res. Methodol.*, vol. 21, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/s12874-021-01272-3.

[64] D. Xu, J. Q. Sheng, P. J. H. Hu, T. S. Huang, and C. C. Hsu, "A deep learning–based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2260–2272, Jun. 2021, doi: 10.1109/JBHI.2020.3033323.

[65] S. Kumar, M. K. Pandey, A. Nath, and K. Subbiah, "Performance analysis of ensemble supervised machine learning algorithms for missing value imputation," in *Proc. 2nd Int. Conf. Comput. Intell. Netw. (CINE)*, Jan. 2016, pp. 160–165, doi: 10.1109/CINE.2016.35.

[66] B. Conroy, L. Eshelman, C. Potes, and M. Xu-Wilson, "A dynamic ensemble approach to robust classification in the presence of missing data," *Mach. Learn.*, vol. 102, no. 3, pp. 443–463, Mar. 2016, doi: 10.1007/s10994-015-5530-z.

[67] M. N. M. Salleh and N. A. Samat, "FCMPSO: An imputation for missing data features in heart disease classification," in *Proc. IOP Conf., Mater. Sci. Eng.*, 2017, vol. 226, no. 1, Art. no. 012102, doi: 10.1088/1757-899X/226/1/012102.

[68] Y. Yan, Y. Wu, X. Du, and Y. Zhang, "Incomplete data ensemble classification using imputation-revision framework with local spatial neighborhood information," *Appl. Soft Comput.*, vol. 99, Feb. 2021, Art. no. 106905, doi: 10.1016/j.asoc.2020.106905.

[69] Z. Hu and D. Du, "A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction," *PLoS ONE*, vol. 15, no. 9, 2020, Art. no. e0237724, doi: 10.1371/journal.pone.0237724.

[70] H. A. Khorshidi, M. Kirley, and U. Aickelin, "Machine learning with incomplete datasets using multi-objective optimization models," arXiv, 2020.

[71] L. R. Galvão and L. H. C. Merschmann, "HSIM: A supervised imputation method for hierarchical classification scenario," *Lecture Notes in Computer Science (Including Subseries Lecture Notes Artificial Intelligent Lecture Notes Bioinformatics)*, vol. 9956. 2016, pp. 134–148, doi: 10.1007/978-3-319-46307-0_9.

[72] S. Oehmcke, O. Zielinski, and O. Kramer, "KNN ensembles with penalized DTW for multivariate time series imputation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2774–2781, doi: 10.1109/IJCNN.2016.7727549.

[73] F. Saitoh, "An ensemble model of self-organizing maps for imputation of missing values," in *Proc. IEEE 9th Int. Workshop Comput. Intell. Appl. (IWCIA)*, Nov. 2016, pp. 9–14, doi: 10.1109/IWCIA.2016.7805741.

[74] O. Kadri, L. H. Mouss, and A. Abdelhadi, "Fault diagnosis for a milk pasteurisation plant with missing data," *Int. J. Qual. Eng. Technol.*, vol. 6, no. 3, pp. 123–136, 2017, doi: 10.1504/IJQET.2017.088858.

[75] O. Elezaj, S. Yildirim, and E. Kalemi, "Data-driven machine learning approach for predicting missing values in large data sets: A comparison study," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes Artificial Intelligent Lecture Notes Bioinformatics)*, vol. 10710. 2018, pp. 268–285, doi: 10.1007/978-3-319-72926-8_23.

[76] G. Nagarajan and L. D. Dhinesh Babu, "A hybrid of whale optimization and late acceptance Hill climbing based imputation to enhance classification performance in electronic health records," *J. Biomed. Informat.*, vol. 94, Jun. 2019, Art. no. 103190, doi: 10.1016/j.jbi.2019.103190.

[77] T.-T.-H. Phan, "Machine learning for univariate time series imputation," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, Oct. 2020, pp. 1–6, doi: 10.1109/MAPR49794.2020.9237768.

[78] S. Nikfalazar, C.-H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowl. Inf. Syst.*, vol. 62, no. 6, pp. 2419–2437, Jun. 2020, doi: 10.1007/s10115-019-01427-1.

[79] S. Daberdaku, E. Tavazzi, and B. Di Camillo, "A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data," *J. Healthcare Informat. Res.*, vol. 4, no. 2, pp. 174–188, Jun. 2020, doi: 10.1007/s41666-020-00069-1.

[80] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020, doi: 10.1109/ACCESS.2020.2994033.

[81] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi, "Similarity-learning information-fusion schemes for missing data imputation," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104805, doi: 10.1016/j.knosys.2019.06.013.

[82] T. Su, Y. Shi, J. Yu, C. Yue, and F. Zhou, "Nonlinear compensation algorithm for multidimensional temporal data: A missing value imputation for the power grid applications," *Knowl.-Based Syst.*, vol. 215, Mar. 2021, Art. no. 106743, doi: 10.1016/j.knosys.2021.106743.

[83] Q. Yuan, M. Longo, A. W. Thornton, N. B. McKeown, B. Comesaña-Gándara, J. C. Jansen, and K. E. Jelfs, "Imputation of missing gas permeability data for polymer membranes using machine learning," *J. Membrane Sci.*, vol. 627, Jun. 2021, Art. no. 119207, doi: 10.1016/j.memsci.2021.119207.

[84] E.-L. Silva-Ramírez and J.-F. Cabrera-Sánchez, "Co-active neuro-fuzzy inference system model as single imputation approach for non-monotone pattern of missing data," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 8981–9004, Aug. 2021, doi: 10.1007/s00521-020-05661-5.

[85] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data," *Soft Comput.*, vol. 25, no. 8, pp. 5993–6012, Apr. 2021, doi: 10.1007/s00500-021-05590-y.

[86] W. Lan, X. Chen, T. Zou, and C. L. Tsai, "Imputations for high missing rate data in covariates via semi-supervised learning approach," *J. Bus. Econ. Statist.*, to be published, doi: 10.1080/07350015.2021.1922120.

[87] J. Huang, J. W. Keung, F. Sarro, Y.-F. Li, Y. T. Yu, W. K. Chan, and H. Sun, "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study," *J. Syst. Softw.*, vol. 132, pp. 226–252, Oct. 2017, doi: 10.1016/j.jss.2017.07.012.

[88] D. Zeng, D. Xie, R. Liu, and X. Li, "Missing value imputation methods for TCM medical data and its effect in the classifier accuracy," in *Proc. IEEE 19th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2017, pp. 77–80, doi: 10.1109/HealthCom.2017.8210844.

[89] T. Mahboob, A. Ijaz, A. Shahzad, and M. Kalsoom, "Handling missing values in chronic kidney disease datasets using KNN, K-means and K-medoids algorithms," in *Proc. 12th Int. Conf. Open Source Syst. Technol. (ICOSST)*, Dec. 2018, pp. 76–81, doi: 10.1109/ICOSST.2018.8632179.

[90] E. Tavazzi, S. Daberdaku, R. Vasta, A. Calvo, A. Chiò, and B. Di Camillo, "Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data with an adaptive K-nearest neighbours approach," *BMC Med. Informat. Decis. Making*, vol. 20, no. S5, pp. 1–23, Aug. 2020, doi: 10.1186/s12911-020-01166-2.

[91] B. Ramosaj and M. Pauly, "Predicting missing values: A comparative study on non-parametric approaches for imputation," *Comput. Statist.*, vol. 34, no. 4, pp. 1741–1764, Dec. 2019, doi: 10.1007/s00180-019-00900-3.

[92] S. Zhang, L. Gong, Q. Zeng, W. Li, F. Xiao, and J. Lei, "Imputation of GPS coordinate time series using MissForest," *Remote Sens.*, vol. 13, no. 12, p. 2312, 2021, doi: 10.3390/rs13122312.

[93] J. Ke, S. Zhang, H. Yang, and X. Chen, "PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data," *Transportmetrica A, Transp. Sci.*, vol. 15, no. 2, pp. 872–895, Nov. 2019, doi: 10.1080/23249935.2018.1542414.

[94] I. E. W. Rachmawan and A. R. Barakbah, "Optimization of missing value imputation using reinforcement programming," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2015, pp. 128–133, doi: 10.1109/ELECSYM.2015.7380828.

[95] M. Zhu and H. Shi, "A novel support vector machine algorithm for missing data," in *Proc. 2nd Int. Conf. Innov. Artif. Intell. (ICIAI)*, 2018, pp. 48–53, doi: 10.1145/3194206.3194214.

[96] A. Ngueilbaye, H. Wang, D. A. Mahamat, and S. B. Junaidu, "Modulo 9 model-based learning for missing data imputation," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107167, doi: 10.1016/j.asoc.2021.107167.

[97] Y. Liang, L. Bi, and X. Su, "Missing data recovery in large-scale, sparse datacenter traces: An Alibaba case study," in *Proc. 19th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGRID)*, May 2019, pp. 251–261, doi: 10.1109/CCGRID.2019.00039.

[98] A. Nekouie and M. H. Moattar, "Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 31, no. 3, pp. 287–294, Jul. 2019, doi: 10.1016/j.jksuci.2018.01.006.

[99] B. D. Chivers, J. Wallbank, S. J. Cole, O. Sebek, S. Stanley, M. Fry, and G. Leontidis, "Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach," *J. Hydrol.*, vol. 588, Sep. 2020, Art. no. 125126, doi: 10.1016/j.jhydrol.2020.125126.

[100] P. Körner, R. Kronenberg, S. Genzel, and C. Bernhofer, "Introducing gradient boosting as a universal gap filling tool for meteorological time series," *Meteorologische Zeitschrift*, vol. 27, no. 5, pp. 369–376, Dec. 2018, doi: 10.1127/metz/2018/0908.

[101] G. Madhu, B. L. Bharadwaj, G. Nagachandrika, and K. S. Vardhan, "A novel algorithm for missing data imputation on machine learning," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 173–177, doi: 10.1109/ICSSIT46314.2019.8987895.

[102] X. Zhang, C. Yan, C. Gao, B. Malin, and Y. Chen, "XGBoost imputation for time series data," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2019, pp. 1–3, doi: 10.1109/ICHI.2019.8904666.

[103] X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, "Predicting missing values in medical data via XGBoost regression," *J. Healthcare Informat. Res.*, vol. 4, no. 4, pp. 383–394, Dec. 2020, doi: 10.1007/s41666-020-00077-1.

[104] D. A. Rusdah and H. Murfi, "XGBoost in handling missing values for life insurance risk prediction," *Social Netw. Appl. Sci.*, vol. 2, no. 8, p. 1336, Aug. 2020, doi: 10.1007/s42452-020-3128-y.

[105] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," in *Engineering for Sustainable Future* (Lecture Notes in Networks and Systems), vol. 101. 2020, pp. 215–227, doi: 10.1007/978-3-030-36841-8_21.

[106] B. K. Beaulieu-Jones and J. H. Moore, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Proc. Pacific Symp. Biocomput.*, 2017, pp. 207–218, doi: 10.1142/9789813207813_0021.

[107] C. Li, "Little's test of missing completely at random," *Stata J.*, vol. 13, no. 4, pp. 795–809, 2013, doi: 10.1177/1536867x1301300407.

[108] M. M. Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, and M. Modat, "Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling," *Med. Image Anal.*, vol. 53, pp. 39–46, Apr. 2019, doi: 10.1016/j.media.2019.01.004.

[109] M. Vilardell, M. Buxó, R. Clèries, and J. M. Martínez, "Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival," *Artif. Intell. Med.*, vol. 107, Jul. 2020, Art. no. 101875, doi: 10.1016/j.artmed.2020.101875.

[110] H. Wang, Y. Chen, B. Shen, D. Wu, and X. Ban, "Generative adversarial networks imputation for high rate missing values," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jul. 2018, pp. 586–590, doi: 10.1109/Cybermatics_2018.2018.00121.

[111] J.-S. Kim, X. Gao, and A. Rzhetsky, "RIDDLE: Race and ethnicity imputation from disease history with deep LEarning," 2017, *arXiv:1707.01623*.

[112] J. Gupta, S. Paul, and A. Ghosh, *A Novel Transfer Learning-Based Missing Value Imputation on Discipline Diverse Real Test Datasets—A Comparative Study With Different Machine Learning Algorithms*, vol. 814. Singapore: Springer 2019.

[113] A. Marshall, D. G. Altman, P. Royston, and R. L. Holder, "Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study," *BMC Med. Res. Methodol.*, vol. 10, no. 1, Dec. 2010, doi: 10.1186/1471-2288-10-7.

MUSTAFA ALABADLA received the B.Sc. degree in information technology from the College of Science and Technology, Palestine, in 2011, and the M.Sc. degree in informatics from the Universiti Sains Malasyia, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. He is also a Software Engineer with an extensive experience in developing cross-platform mobile apps and UI design. He worked in several IT fields, including web development, database administration, computer networks, digital marketing, and software engineering. His research interests include ML, database systems, software engineering, and information systems.

FATIMAH SIDI (Member, IEEE) received the Ph.D. degree in management information system from the Universiti Putra Malaysia (UPM), Malaysia, in 2008. She is currently working as an Associate Professor in the discipline of computer science with the Department of Computer Science, Faculty of Computer Science and Information Technology, UPM. Her current research interests include knowledge and information management systems, data and knowledge engineering, database, data warehouse, big data, and data analytics.

ISKANDAR ISHAK received the Bachelor of Information Technology degree from the Universiti Tenaga Nasional, Malaysia, the Master of Technology degree in information technology from the Royal Melbourne Institute of Technology, Australia, and the Ph.D. degree in computer science from the Universiti Teknologi Malaysia. His research interests include database systems, big data, and data analytics.

HAMIDAH IBRAHIM (Member, IEEE) received the Ph.D. degree in computer science from the University of Wales, Cardiff, U.K., in 1998. She is currently a Full Professor with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). Her current research interests include databases (distributed, parallel, mobile, biomedical, and XML) focusing on issues related to integrity maintenance/checking, ontology/schema/data integration, ontology/schema/data.

LILLY SURIANI AFFENDEY received the Bachelor of Computer Science degree from the Universiti Pertanian Malaysia, in 1991, the M.Sc. degree in computing from the University of Bradford, U.K., in 1994, and the Ph.D. degree from the Universiti Putra Malaysia, in 2007. She is currently an Associate Professor with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She teaches database application development, database systems, business analytics, and big data analytics. She has recently attended trainings on data science, RapidMiner, Talend, and Hadoop. Her current research interests include multimedia databases, video content-based retrieval, data science, and big data analytics.

ZAFIENAS CHE ANI received the Bachelor of Computer Science degree from the Universiti Putra Malaysia, in 2014, and the Master of Computer Science degree in artificial intelligence from the University of Malaya, in 2017. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She is also a System Analyst at the National Defence University of Malaysia. Her current research interests include ML and data science.

MARZANAH A. JABAR received the Ph.D. degree in management information system from the Universiti Putra Malaysia, Malaysia. She has over 20 years of experience as a System Analyst in the area of enterprise system development and has been appointed as a consultant to several software development projects in UPM and other agencies. She is currently a Lecturer with the Department of Software Engineering and Information System. She is the Principal Investigator for 35 research grants valued at RM2 millions, consultation work worth at RM300,000 and has published in more than 200 articles in journals, conferences proceedings, seminars, and technical reports. Additionally, she has 53 copyrights and two patents to her name. She has also successfully commercialized one product from her own research. Her current research interests include software engineering, knowledge management, information management systems, and enterprise software development.

UMAR ALI BUKAR received the B.Sc. degree in business information technology with concentration in e-commerce research and strategy from Greenwich University, U.K., and the M.Sc. degree in computer network management from Middlesex University, Dubai. He is currently pursuing the Ph.D. degree with the Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His contributions have been published in prestigious peer-reviewed journals and international conferences. His IT career has included work as several niche projects, with responsibilities ranging from teaching, research, and analysis. His research interests include crisis informatics, data analytic, ML, and the use of quantitative methods in information systems research.

NAVIN KUMAR DEVARAJ received the Master of Medicine degree in family medicine from University of Malaya. He is currently a Family Medicine Specialist and a Medical Lecturer with the Faculty of Medicine and Health Sciences, Universiti Putra Malaysia (UPM). He has been a Physician for more than 15 years. He has published more than 40 articles in reputable journals worldwide and has been actively involved in both academic and community level events, especially as an Advisor of Asian Medical Student Association, UPM Chapter, and a member of the Integrative and Accident Prevention committees of Malaysian Medical Association. He is a Keen Teacher, a Researcher, and a Keen Supporter of ethics in teaching and learning. His research interests include hypertension, anti-smoking, hyperlipidaemia, and men's health.

AHMAD SOBRI MUDA received the Medical degree in medicine from the Universiti Kebangsaan Malaysia, followed by specialist training in radiology from the Universiti Sains Malaysia. He is currently a Professor of radiology with the Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Malaysia. His main clinical practice in Hospital Pengajar UPM, with focus on neuroradiology, neurointervention, and stroke. He has few patents, industrial designs related to medical technology and one of the Founding Member of padimedical system. His research interests include neuroimaging, medical technology, ML, and stroke.

ANAS THAREK received the Master of Radiology degree from the Universiti Putra Malaysia (UPM). He is currently a Radiologist and a Medical Lecturer at the Department of Radiology, Faculty of Medicine and Health Sciences, UPM. He is one of the developers for Padimedical system, which is use to manage DICOM medical imaging in hospital. His research interests include big data, ML, and medical informatics.

NORITAH OMAR is currently an Associate Professor with the English Department, Universiti Putra Malaysia, where she teaches literary theory, Malaysian literature in English, and feminism and social change. She specializes in narrative and critical ethnography, and gender studies. She has published articles and book chapters on critical literacy, gender and Islam, Malaysian and Singapore Literature, Islam and modern Malay literature, Islam and contemporary popular culture, and issues in postgraduate studies.

M. IZHAM MOHD JAYA (Member, IEEE) received the Ph.D. degree in database management system from the Universiti Putra Malaysia (UPM), Malaysia, in 2018. He is currently working as a Senior Lecturer at the Department of Software Engineering, Faculty of Computing, Universiti Malaysia Pahang (UMP), Malaysia. His current research interests include data quality, data management, and ML.