Missing Value Imputation Designs and Methods of Nature-Inspired Metaheuristic Techniques: A Systematic Review

PO CHAN CHIU1,2, ALI SELAMAT1,3,4, (Member, IEEE), ONDREJ KREJCAR3,4, KING KUOK KUOK5, SITI DIANAH ABDUL BUJANG3 AND HAMIDO FUJITA3,6,7,8

1 School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia & MagicX (Media and Games Center of Excellence), Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia
2 Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300, Kota Samarahan, Sarawak, Malaysia
3 Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100, Kuala Lumpur, Malaysia
4 Faculty of Informatics and Management, University of Hradec Kralove, Rokitanského 62, 500 03 Hradec Kralove, Czech Republic
5 Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, 93350 Kuching, Sarawak, Malaysia
6 Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain
7 i-SOMET incorporated Association, Morioka, Japan
8 Regional Research Center, Iwate Prefectural University, Iwate, Japan

Corresponding authors: Ali Selamat (aselamat@utm.my) and Po Chan Chiu (pcchiu@unimas.my)

ABSTRACT Missing values are highly undesirable in real-world datasets. The missing values should be estimated and treated during the preprocessing stage. With the expansion of nature-inspired metaheuristic techniques, interest in missing value imputation (MVI) has increased. The main goal of this literature is to identify and review the existing research on missing value imputation (MVI) in terms of nature-inspired metaheuristic approaches, dataset designs, missingness mechanisms, and missing rates, as well as the most used evaluation metrics between 2011 and 2021. This study ultimately gives insight into how the MVI plan can be incorporated into the experimental design. Using the systematic literature review (SLR) guidelines designed by Kitchenham, this study utilizes renowned scientific databases to retrieve and analyze all relevant articles during the search process. A total of 48 related articles from 2011 to 2021 was selected to assess the review questions. This review indicated that the synthetic missing dataset is the most popular baseline test dataset to evaluate the effectiveness of the imputation strategy. The study revealed that missing at random (MAR) is the most common proposed missing mechanism in the datasets. This review also indicated that the hybridizations of metaheuristics with clustering or neural networks are popular among researchers. The superior performance of the hybrid approaches is significantly attributed to the power of optimized learning in MVI models. In addition, perspectives, challenges, and opportunities in MVI are also addressed in this literature. The outcome of this review serves as a toolkit for the researchers to develop effective MVI models.

INDEX TERMS Missing Value, Missing Data, Imputation, Incomplete Dataset, Metaheuristic, Systematic Review.

I. INTRODUCTION

Data quality in machine learning has been intensively studied over the past decades. One of the data quality issues is missing values. Missing values can be defined as portions of the data that are either incomplete or absent in the dataset. The presence of missing values in the dataset diminishes data quality, reduces the power of data analysis, and induces bias in data science applications. Hence, dealing with incomplete information is critical for most data mining and machine learning techniques [1].

Numerous studies have been successfully conducted to address the issue of missing values. Little and Rubin [2]
classified missing values into three mechanisms, missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR). In the case of MCAR, the probability of missing values is independent. In the type of MAR, the probability of data incomplete is not related to the missing value; instead, it is related to the part of the observed data. In the MNAR case, missing values are dependent on the missing variable, in which the incomplete values are associated with unmeasured events. Any missing value estimation technique could be applied due to the absence of bias in the data.

Furthermore, the missing value pattern explains how the data is missing in different ways. Univariate missing value pattern occurs when only one variable is missing. Data is missing monotone if the missing values follow a pattern. On the other hand, data is missing arbitrarily if the data is missing without a clear pattern.

Moreover, the percentage of missing values impacting the data quality. However, the existing literature does not have a standard cutoff for the acceptable proportion of missing values in a dataset for quality data analysis. For example, Bormann [3] suggested that 10% missing precipitation values of the calendar days are the threshold for removing the whole winter observations from the analysis. In contrast, Tatar et al. [4] stated that a threshold of 50% missing features was excluded from the prediction of low salinity waterflooding, while imputation of mean value was applied for missing features below the missing threshold.

Equipment failure is a major cause of high missing rates. Eliminating high missing rates from the observations diminishes the representativeness of the samples. However, missing values can be higher than 50% in real-world scenarios. Therefore, missing value imputation (MVI) is used to address the problem of missing values. MVI is a procedure that is used to fill in missing values with substitutes [5]. Over the past decades, various machine learning techniques have been proposed to deal with incomplete datasets for different domain problems, such as medical [6], hydrology [7], [8], and transportation [9].

Consequently, a number of literature [10]–[12] discusses recent machine learning-based imputation techniques in solving incomplete dataset problems. Nevertheless, with respect to MVI of nature-inspired metaheuristic techniques, the literature receives limited attention. Therefore, the purpose of this literature is to review recent MVI designs of metaheuristic techniques used for handling and optimizing missing value imputation. This SLR follows the guidelines established by Kitchenham and Charters [13], thereby providing significant insights for researchers working in the MVI domain.

The contributions of this literature are:
1) A comprehensive systematic literature review on the existing MVI designs for metaheuristic approaches is presented, together with experimental design, dataset design, missingness mechanisms, missing rates, and evaluation metrics.
2) A guide to address, manage, and report MVI studies is introduced. This SLR serves as a toolkit for the researchers to come up with solutions for challenges in implementing effective missing value imputation.

This research is organized as follows: Section II presents the SLR methodologies, whereas Section III summarizes the SLR findings. Section IV discusses the research trends and potential opportunities in MVI. Section V highlights the challenges, and finally, the conclusion is presented in Section VI.

II. RESEARCH AND REVIEW METHOD
This section describes the systematic approach for reviewing recent articles on metaheuristic-based MVI techniques by adopting Kitchenham's SLR standards. This SLR is inspected, analyzed, and evaluated according to the research questions and review protocols. Each phase of this SLR is explained in the following sections.

A. PLANNING THE REVIEW
This section outlines the review plan needed to undertake the SLR, which includes formulating research questions in accordance with the review’s primary objective, defining a search strategy, and designing a comprehensive review protocol.

1) RESEARCH QUESTIONS
This review aims to study the existing literature on metaheuristic designs and methods for optimizing and solving missing value problems. The following Research Questions (RQs) for this literature are formulated to accomplish this aim, as indicated in Table 1.

In the past ten years, several novel imputation techniques have been proposed. This SLR aims to identify the differences among the methods to enrich the understanding of MVI methods, which can be taken as the basis for planning and developing a new imputation model. RQ1 is formulated to provide an overview of state-of-the-art metaheuristic techniques used to handle and optimize missing value imputation. Meanwhile, RQ2 is defined to explore the experimental designs of imputation and understand what factors affect the MVI design. RQ3 is outlined to understand what metrics are commonly used when evaluating the missing value imputation method.

2) SEARCH STRATEGY
The search strategy begins with selecting relevant databases (IEEEexplore, ScienceDirect, Scopus, and other electronic databases) to track scientific papers that address research topics published in linked journals, conferences, and book chapters. The search string used to retrieve articles from the scientific databases is described as follows:
3) INCLUSION AND EXCLUSION CRITERIA
A list of inclusion and exclusion criteria was constructed in this literature, as shown in Table 2. The inclusion and exclusion criteria are used as one of the review protocols to narrow the relevant studies to the most pertinent ones during the article review process.

4) QUALITY ASSESSMENT CRITERIA
Another review protocol is the quality assessment criteria. The quality assessment criteria are crucial to determining the selected articles’ quality. A quality assessment criteria constructed based on Kitchenham and Charters [13], Genc-Nayebi et al. [14], and Yang et al. [15] is presented in Table 3. The quality assessment is based on the response of “Yes,” “No,” and “Partial applicable,” abbreviated as “Y,” “N,” and “P,” respectively.

B. CONDUCTING THE REVIEW
The article selection was carried out by applying the mentioned search string. Initially, our search string found 758 publications from different databases between 2011 and 2021. The search results were then narrowed down to manually reviewing all the articles’ titles and abstracts, resulting from a total of 644 articles. Next, the potential articles were filtered according to the RQs, which yielded 181 articles. Further filtering was applied by removing irrelevant studies according to the detailed inclusion and exclusion criteria, as shown in Table 2. Additionally, the quality assessment was conducted, and we chose articles that affirmatively respond to the nine quality assessment criteria listed in Table 3. The findings indicated that most selected articles satisfied all the quality assessment criteria. On final selection, a total of 48 articles fulfilled all the inclusion and quality assessment criteria used in this literature. The article selection processes are summarized and illustrated in Figure 1.

---

**TABLE 1. List of research questions.**

| No | Research Questions | Motivation |
|----|--------------------|------------|
| 1  | What are the existing metaheuristic techniques used for handling and optimizing missing value imputation? | Identify the state-of-the-art metaheuristic techniques used for solving missing value problems. |
| 2  | What are the factors affecting missing value imputation design? | Identify the experimental designs used for imputation. |
| 3  | What are the commonly used metrics to evaluate the performance of the missing value imputation? | Identify the most common metric used to assess missing value imputation performance. |

**TABLE 2. The inclusion and exclusion criteria.**

| Inclusion criteria | Exclusion criteria |
|--------------------|--------------------|
| Articles that are published from 2011 till 2021 | Articles that are published before 2011. |
| All related articles that match with the research questions | Articles that do not address the research questions. |
| All articles published in English language | Articles that are published not in English language |

**TABLE 3. Quality assessment criteria and results of selected articles.**

| No | Quality assessment criteria | Y | N | P |
|----|-----------------------------|---|---|---|
| 1  | Are the objectives of the research clear and relevant? | 48 | 0 | 0 |
| 2  | Is the proposed technique described in detail? | 47 | 0 | 1 |
| 3  | Is the research design appropriate to address the aims of the study? | 48 | 0 | 0 |
| 4  | Is the incomplete dataset adequately described? | 48 | 0 | 0 |
| 5  | Is the missing mechanism described in detail? | 38 | 10 | 0 |
| 6  | Is the missing rate clearly defined? | 41 | 7 | 0 |
| 7  | Are the evaluation metrics used in the research well documented? | 48 | 0 | 0 |
| 8  | Are the findings of the research reliable? | 39 | 0 | 9 |
| 9  | Is the data analysis sufficiently rigorous? | 37 | 0 | 11 |

**FIGURE 1. The process of article selection.**

---

**III. RESEARCH FINDINGS**
This section presents and discusses the findings from the literature review conducted in response to the RQs identified in Section II. This section is divided into three subsections: the first illustrates state-of-the-art metaheuristic techniques...
for managing and optimizing MVI. The second subsection discusses the experimental designs and factors affecting MVI design. Finally, the third subsection explores the various evaluation metrics that are used to evaluate the performance of MVI.

A. SUMMARY OF METAHEURISTIC TECHNIQUES USED IN MANAGING AND OPTIMIZING MISSING VALUE IMPUTATION

This subsection mainly focuses on RQ1, which identifies metaheuristic techniques for handling and optimizing MVI. Figure 2 indicates the trend of publications over ten years. The graph illustrates the popularity of metaheuristic techniques in MVI research over time. As can be seen, studies on metaheuristic-based MVI have experienced continuous growth since 2011 and show an emerging trend in MVI research. The growth is apparently due to the explosion of data science research involving high-quality data, which raised researchers’ awareness of the importance of imputation.

![Figure 2. Year-wise distribution of publications relevant to studies.](image)

Next, we summarize the metaheuristic techniques employed in handling MVI and highlight their primary benefits. We have categorized the metaheuristic technologies into three categories. The first category is a single objective approach, followed by multi-objective and hybrid approaches as the second and third categories. The taxonomy of metaheuristic approaches in handling and optimizing MVI is shown in Figure 3.

From the literature, genetic algorithm (GA) has become one of the most widely used metaheuristic approaches in MVI tasks. Figueroa García et al. [16] used GA imputation to estimate missing values by minimizing an error function derived from covariance matrix and means vector, while Lobato et al. [17] improved GA imputation for the incomplete multi-attribute dataset. Recently, Awawdeh et al. [18] performed imputation and feature selection simultaneously. GA was used to determine the most optimal features, while mean and mode imputation were used for filling missing numeric and categorical features, respectively. The advantage of this approach is that it is more tolerant of bias in MAR and NMAR missingness types. In another study, Sivapragasam et al. [19] utilized mathematical models in genetic programming (GP) to reconstruct missing time series rainfall data. In [20], PSO imputation was proposed to infill missing gene expressions. The advantages of this approach are it is simple and easy to implement. However, the performance of the PSO imputation cannot be generalized as it is only compared with conventional imputers such as K-nearest neighbor (KNN) and row averaging imputation at missing rates of 5%, 8%, and 10%.

For multi-objective metaheuristic approaches, Lobato et al. [21] analyzed incomplete instances and modeling task information using multi-objective GA (MOGA-II) based non-dominated sorting genetic algorithm-II (NSGA-II) to infill mixed-attribute datasets. Both objective functions of root mean square error (RMSE) and classification accuracy significantly improved the imputation performances for incomplete numeric and nominal features. On the other hand, recent work by Khorshidi et al. [22] proposed two objective functions of cluster validity function and correlation function to enhance the existing NSGA-II. The advantages of this approach are that it is robust and able to handle online imputation and classification simultaneously for MAR missingness type. The proposed multi-objective particle swarm optimization (MOPSO) approach in [23] determined the optimal imputation algorithm based on the MCAR, MAR, and MNAR missingness mechanisms, in which the fitness function adapted according to sensitivity and specificity. The proposed MOPSO improved the imputation accuracy by 16.52% than the delete missing, mean, expectation-maximization, multivariate imputation by chained equations (MICE), and missForest imputation approaches. However, the shortcomings of this approach are that it is slow, and the imputation model is more dependent on variables than on records.

Several new methods have been proposed to improve imputation accuracy that combines metaheuristic methods with other techniques such as Bayesian, clustering, probabilistic, and neural network. Furthermore, most studies adopted hybrid approaches to address missing value issues. As for the Bayesian category, several studies [24]–[27] have explored the idea of infilling MVI using the combination of metaheuristic and Bayesian algorithms. The advantage of incorporating bayesian fitness is improved the solution's optimality. In [28], Nekouie and Moattar improved imputation performance using bayesian, tensor, and chaotic PSO. The approach significantly reduced 4% error than the tensor method for missing numerical values and class imbalance problems.

On the other hand, some researchers combined probabilities and metaheuristics approaches to estimate missing values [29]–[33]. KNN imputation was used to infill missing values based on neighbors' data and optimized by GA [29] and PSO [30]. Recently, Nagarajan and Dhinesh
Babu [31] proposed a feature weighting approach that combined an improved local search and whale optimization algorithm (WOA). The advantage of this approach is that the hybrid learned various k of nearest neighbor for different testing values by examining the correlation matrix between the training and testing datasets. Moreover, the WOA avoided local optima and converged to a better solution in final iterations. The findings indicated that missing values were predicted more precisely and improved classification performance in electronic health records. However, this approach is inefficient in large datasets with high dimensional features. Meanwhile, Krishna and Ravi [32] utilized a covariance matrix to reduce the error function of PSO. The approach achieved better classification accuracy than the hybrid K-means and multilayer perceptron (MLP), producing comparable results for regression tasks. In the time series problem, a combination of inverse distance weight (IDW), tolerance rough set (TR), and PSO [33] was proposed to determine the optimal influence factor value for each recognized data point in the neighboring group, thereby reducing the error rate of the imputed time series data.

As for the clustering category, several researchers employed the clustering method with soft computing. In [34], Veroneze et al. proposed a combination of bi-clustering and ant colony optimization (ACO) to deal with missing data problems. The introduction of a bi-clustering strategy and optimal parameter selection in this approach enhanced the imputation quality for the missing gene expression datasets; however, the impact of long execution times increased the computational cost of this approach.

The works of [35] and [36] specified the use of Fuzzy C-means (FCM) with GA by generating a matrix-based data structure and optimizing it through a GA parameter optimization process to improve the accuracy of missing value estimation. Meanwhile, Aydilek and Arslan [37] demonstrated that combining an optimized clustering process with support vector training improved imputation performance. However, higher proportions of 25% missing data were not considered in the study. Then, Khotimah and Pramudita [38] implemented a self-organizing map (SOM) imputation with GA. The selection of SOM weights using GA with elite chromosomes determined the shortest distance between the data and the cluster centroid, resulting in a more accurate solution for incomplete data estimation.

In FCM imputation with the PSO method [39]–[42], the missing values can be estimated from the observed data with different optimized weights to improve data quality. Recent work by Hu et al. [43] presented missing values in hybrid numeric and granular forms. It used information granularities to construct granular fuzzy models (GFM), while PSO optimized the optimal allocation of information granularities. The advantage of this approach is that the established granular models improved numerical value prediction accuracy by extracting the essential target information from incomplete data. On the other hand, Gautam and Ravi [44] implemented data imputation via a two-stage learning strategy: the first stage was based on local learning in particle swarm optimization-evolving clustering method (PSO-ECM), and the second stage was based on global approximation in auto-associative extreme learning machine (AEELM). Another approach is the ELM+PSO+FCM proposed by Sun et al. [45], which resulted in effective data imputation for byproduct gas flow data. These studies [43]–[45] demonstrated a positive impact on MVI accuracy, but the imputation results were only examined at missing rates under 50%.

To provide greater accuracy in predicting numerical and nominal missing values, the recent work in [46] extended the existing PSO imputation approach by incorporating ontology and K-means, where ontology eliminated irrelevant data, and K-means accelerated PSO convergence. In addition to PSO imputation, a fruit fly optimization algorithm (FOA) has been proposed by [47] in solving missing time series values. First, SOM was used to cluster the time series and obtain a similarity matrix for the incomplete series. Then, this approach employed a cross-validation procedure and FOA strategy to determine the optimal parameter in the least-squares support vector machine (LSSVM) for building an optimal imputation model. In addition, Tran et al. [48] proposed an approach for classifying missing values that integrated imputation, clustering, and feature selection. The proposed clustering minimized the number of instances used by imputation, whereas differential evolution (DE) extracted relevant features of the training data. However, removing instances may result in data loss, and performing feature selection after initial imputation can be time-consuming, particularly when dealing with high-dimensional data.

Duma et al. [50] proposed a hybrid multi-layered artificial immune system and GA to fill in missing values for insurance datasets. In [49], the authors demonstrated that using random forest (RF) and GA-selected predictors to estimate missing forest inventory variables with data from target and auxiliary stands significantly reduced model bias. Other than that, the proposed hybrid GA and asexual reproduction optimization (ARO) approach outperformed the mean and original GA imputation approaches by incorporating ARO imputation and GA optimization [51].

A published work in [52] recently improved the existing GP algorithm by designing a mixed tree-vector representation that can be used for selection and symbolic regression on missing data. The imputation performance was improved for medium-sized datasets; nevertheless, it was less significant for datasets with relatively small instances (< 300), a large number of instances (> 8191), or below missing rates of 2%. In addition, this imputer model also has the drawback of requiring a large volume of data for training.
FIGURE 3. Taxonomy of metaheuristic techniques based on missing value imputation.
### TABLE 4. State-of-the-art metaheuristic techniques for handling and optimizing missing value imputation.

| Category       | Technique    | Description                                                                 | Strengths                                                                 | Studies               |
|----------------|--------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------|-----------------------|
| Single objective | GA          | Employed minimization of an error function derived from covariance matrix and means vector of related data to estimate missing values. | The proposed approach enhanced imputation for missing multivariate data   | [16]                  |
|                |             | GA imputation to find the best estimate values for filling missing values in a multi-attribute dataset. | This approach improved the classification accuracy for mixed variable types. | [17]                  |
|                |             | Handling missing value imputation and feature selections simultaneously.     | This approach can minimize bias when handling MAR and NMAR missing data types. | [18]                  |
|                | GP          | GP incorporated with mathematical models such sin, cos, exp, and log, to predict missing monthly rainfall data. | The approach was able to handle the nonlinear relationship of rainfall data. | [19]                  |
|                | PSO         | PSO based imputation for missing gene expressions.                          | Simple and easy to implement.                                             | [20]                  |
| Multi objective | MOGA-II      | Employed multi objective GA based on the NSGA-II, which can handle mixed-attribute datasets and incorporated information from incomplete instances and modeling tasks. | Significantly improved imputation performances and has higher statistical ranking than the compared methods in both objective functions studied (RMSE and classifier accuracy). | [21]                  |
|                |             | Proposed multi objective optimization model with two objective functions (cluster validity function and correlation function) for imputation and model selection. | Concurrently performed online imputation and classification. It is robust and works well in various situations. | [22]                  |
|                | MOPSO       | The approach proposed the optimal imputation algorithm based on missing data type. | The imputation accuracy was improved by 16.52% than the compared methods. | [23]                  |
| Hybrid         | Bayesian ACO+ | ACO was hybridized with Bayesian principles for imputing the missing values with MAR mechanism. | The proposed approach performed better in estimating discrete and continuous missing values in large datasets under MAR mechanism, compared to multiple imputation, expectation maximization and kernel imputations. | [24]                  |
|                |             | An average value of mean imputation, distance imputation, and random imputation was used to estimate the missing value. Further, bayesian optimization was integrated into the ACO model. | Bayesian optimization employed posterior and prior probability values to evaluate the fitness function of the ACO. This approach successfully solved the discrete value imputation problems. | [25]                  |
|                |             | Hybridization of bayesian min-max and ACO algorithm. The bayesian fitness, which was incorporated into the proposed model, improved the optimality of the solution. | This approach outperformed the competitive imputation models at different percentages of missing rates, ranging from 5% to 50%. | [26], [27]            |
|                |             | Bayesian networks were used to estimate initial missing values. Additionally, the CRAPSO was used for sample generation to deal with tensor data insufficiency. Finally, a modified tensor factorization approach was used for estimating final missing values. | In the presence of missing numerical values and class imbalance, the proposed approach outperformed the compared methods for missing data estimation. | [28]                  |
| Method                | Description                                                                                                                                                                                                 | Improvement                                                                                   |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Probabilistic        | **GA+KNN** Handle missing value imputation using a genetic algorithm optimized KNN algorithm.                                                                                                             | This approach can identify the optimal value of k and weight each attribute in the dataset.   |
| GMSA+MP              | WKN was used to select neighbors’ data for missing data estimation, while GMSA-MPSO was utilized to optimize feature weights.                                                                                 | This approach showed better estimation accuracy for sensor monitoring manufacturing systems than the compared techniques. |
| KNN+LHCAWOA         | Hybridization of an improved local search and WOA with feature weighted nearest neighbor imputation approach for missing health records.                                                                         | This approach improved classification performances using the imputed health datasets.        |
| PSO+covariance       |                                                                                                                                             |                                                                                               |
| IDW+TR+PSO          | TR employed the rough set concept to determine the neighborhood set for each unknown data point. This was followed by a PSO technique to find the optimal influence factor value for each known data point in the neighborhood set. | In comparison to other imputation techniques such as KNN, expectation maximization, and traditional IDW, the proposed system significantly reduced the error rate of the imputed time series results. |
| Clustering           | **ACO+clustering** The nearest neighbor (Euclidean distance) technique was utilized in the pre-imputation stage. The pre-imputed dataset was then replaced by new estimation using bi-clustering and optimal parameter selection strategies in ACO. | The use of bi-clustering strategy and optimal parameter selection in ACO achieved higher imputation quality than KNN and SVD for MCAR and MAR missing mechanisms, despite its higher computational cost. |
| FCM+GA               | A hybrid method that combined FCM imputation method with GA optimization method. This study proposed a matrix-based data structure and GA parameter optimization process to improve the missing data estimation. | This approach was superior to the competitive imputation models.                              |
| FCM+SVR+GA          | This method employed fuzzy C-means clustering data, which combined SVR and GA to handle low proportion of missing data.                                                                                           | The optimized clustering process combined with support vector training improved the imputation performance significantly. |
| GA+SOM              | Clustering-based imputation, in which the model's weights were updated via a chromosome elite search strategy in GA.                                                                                             | Chromosome elite search strategy was more effective and efficient than non-elite search in GA. |
| FCM+PSO             | This hybrid optimization of PSO and FCM employed a fuzzy clustering approach to impute missing values.                                                                                                         | This approach improved traditional clustering imputation by incorporating PSO to find the most optimal values for filling the missing values. |
| GFM+PSO             | This method utilized information granularities to construct granular fuzzy models, and PSO to optimize the allocation of information granularities.                                                             | The established granular models enhanced the prediction accuracy for numerical values.        |
| PSO-ECM+AAELM       | This approach employed two-stage learning; First stage was a local learning in PSO-ECM and second stage was a global approximation in AAELM.                                                              | The optimal parameter selection of ECM by PSO contributed significantly to the good performances of PSO-ECM and PSO-ECM+AAELM. This approach also improved local learning, global optimization, and global learning of the proposed models. |
| ELM+PSO+FCM         | Prefilled missing values were estimated using membership matrix and related cluster centers by linear interpolation and FCM. The clustering size and weighting factor parameters were optimized by an iterative PSO optimization to enhance the accuracy of FCM. The missing value | This approach improved the model's accuracy by imputing missing values in the byproduct gas flow dataset. |
Imputation was further enhanced in ELM by minimizing the Euclidean distance between the estimated values and the missing values.

**PSO+K-means+ontology** model
- Incorporated ontology and K-means in PSO imputation, in which ontology removed irrelevant data and K-means improved PSO convergence speed.
- The use of ontologies and K-means in PSO imputation significantly reduced errors in predicting missing nominal and numerical data. [46]

**SOM+FOA+ LSSVM**
- Optimization techniques were combined with the clustering method to provide sufficient information and an optimal solution.
- Higher imputation accuracy for dealing with missing spatial-temporal values. [47]

**DE + clustering**
- This study proposed a hybrid of DE with clustering and feature selection for classification with missing values.
- By incorporating clustering and feature selection into imputation, the proposed approach achieved higher accuracy at a lower computational time. [48]

| Method       | Algorithm/Technique                                                                 | Description                                                                                           |
|--------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| Random forest| GA+RF                                                                               | This method utilized target and auxiliary stands (off-site samples) data for imputing missing forest inventory variables. |
| **MAIS**     | MAIS+GA                                                                             | A hybrid multi-layered artificial immune system and GA for partial missing value imputation.           |
| **ARO**      | GA+ARO                                                                              | This approach employed ARO to impute missing values for each feature. The output of ARO (best chromosome) will be used as an initialization input for GA. GA iteratively optimized the solution to find the best optimal solution. |
| **Tree vector** | GP+tree vector                                                                     | Improved version of GP, where a mixed tree-vector representation was proposed for performing instance selection, while GP was used for symbolic regression on missing data. |
| **SVM**      | PSO+levy flight+SVM                                                                 | This approach performed well for filling missing creatinine values.                                      |
| **PSO+LSSVM**| PSO+LSSVM                                                                           | This LSSVM model imputed missing data by combining the previous monitoring data from a node and the current monitoring data from a neighboring node. The parameters of imputation were then optimized by PSO. |
| **Wrapper**  | GP+wrapper                                                                          | Proposed to enhance the symbolic regression performance of missing value estimation.                                      |
| **Neural network** | GSO+MLP                                                                           | This method employed three-layer feed forward neural network, in which the weights and thresholds were optimized by GSO during missing traffic flow data imputation. |
| **GA+MLP, SA+MLP, PSO+MLP, RF+MLP** | SC-FITNET                                                                          | The sine cosine algorithm was used to optimize a neural network for imputing missing rainfall data. |

**SC-FITNET**
- The sine cosine algorithm was used to optimize a neural network for imputing missing rainfall data.
- Effectively imputed time series data at different missing rates and outperformed LSTM approach. [59]
In [53], Ismail et al. incorporated levy flight into PSO to improve global exploration of PSO and helped PSO to escape from local optimum. The results indicated that support vector machine (SVM) imputation, optimized by levy flight PSO achieved the lowest error for filling the incomplete creatinine data than KNN, naïve Bayes, and decision tree imputation. Gao et al. also presented a variant of SVM-based imputation that employed LSSVM optimized by PSO to estimate incomplete values for dose rate and sensor rate data. The results revealed that the PSO+LSSVM approach achieved better accuracy than the LSSVM model [54]. Furthermore, Al-Helali et al. [55]-[56] proposed wrapper-specific GP methods to improve imputation accuracy and symbolic regression performances.

The research done in [57] implemented a hybrid GSO and neural network system to perform missing time series data imputation tasks, and the results demonstrated that the approach could accurately predict incomplete traffic flow data for urban arterial streets. The authors [59] proposed a sine cosine algorithm to optimize a function-fitting neural network to impute incomplete rainfall data. A significant advantage of the method is that it outperformed the long short-term memory (LSTM) method in imputing time-series data at various missing rates. Recent work [60] extended the existing sine cosine algorithm by proposing a novel hybrid sine cosine and fitness dependent optimizer (SC-FDO) to approximate missing rainfall data. The introduction of the modified pace-updating position, random weight factor, and conversion parameter strategies significantly improved the searching accuracy of exploratory and exploitative balance in the proposed SC-FDO. The findings revealed that the SC-FDO based MLP trainer yielded higher imputation accuracy for low and high missing rates compared to sine cosine algorithm (SCA) and fitness-dependent optimizer (FDO) based MLP trainer.

On the other hand, Leke et al. [58] investigated the use of hybrid MLP-based auto-associative neural networks with GA, simulated annealing (SA), PSO, and RF in the prediction and classification of missing values. The GA+MLP, SA+MLP, and PSO+MLP algorithms outperformed the RF+MLP algorithm in prediction. However, the RF+MLP algorithm outperformed the GA+MLP, SA+MLP, and PSO+MLP algorithms for classification problems.

In addition to that, Leke et al. explored missing values in high dimensional datasets with the aid of deep learning (DL) and swarm intelligence approaches such as cuckoo search algorithm (CS), firefly algorithm (FA), and bat algorithm. The essential advantage of proposing hybrid models (DL-CS [61] and DL-Bat [62]) is that both models yielded more accurate estimates than the hybrid MLP models in [58] and DL-FA. One of the shortcomings of a deep neural network is that it is time-consuming. As a result, the DL-CS and DL-Bat have higher computational time than the hybrid MLP approaches. The work in [63] further improved the imputer models of [61], [62] by proposing the hybrid DL and gravitational search algorithm (DL-GSA). The DL-GSA [63] outperformed the DL-CS [61] and DL-Bat [62] at higher accuracy and shorter computational time. A relative comparison of metaheuristic techniques for dealing with MVI is presented in Table 4.

B. EXPERIMENTAL DESIGNS

This subsection focuses on the RQ2 that identifies the experimental designs used for imputation. The three aspects to consider when dealing with missing data: dataset characteristics, the missing mechanisms, and missing rates.

1) DATASET CHARACTERISTICS
TABLE 5 summarizes the datasets and their state-of-the-techniques used in the selected articles. From the 133 datasets shown in

TABLE 5, 88% of the datasets are publicly available, while a total of 16 datasets is real-world datasets from industry or agency sources. The findings revealed that the UCI Machine Learning Repository was the most often used dataset over the last ten years, followed by OpenML and Keel. Of all the UCI datasets used here, iris, forest fires, Pima Indian, and wine datasets are the most used datasets. However, the famous databases are on a small scale, containing the number of feature dimensions of less than 15 and the number of instances less than 800.

TABLE 5. Benchmark datasets and their state-of-the-art techniques.

| Dataset source | Dataset                        | No. of studies | Techniques                                      |
|----------------|--------------------------------|----------------|------------------------------------------------|
| Kaggle         | Cancer                         | 1              | KNN+LAHCAWOA [31]                               |
|                | Diabetes                       | 1              |                                                |
|                | Heart                          | 1              |                                                |
|                | Spine                          | 1              |                                                |
| KEEL           | California                     | 1              | GFM+PSO [43]                                    |
|                | Corel                          | 1              |                                                |
|                | Parkinsons                     | 1              |                                                |
|                | Stock                          | 1              |                                                |
|                | Treasury                       | 1              |                                                |
|                | Wankara                        | 1              |                                                |
| MNIST          | Handwritten digits             | 3              | DL-CS [61], DL-BAT [62], DL-GSA [63]            |
| NHLBI          | Framingham heart dataset       | 2              | FCM+PSO [40], [41]                              |
| OpenML         | Bank32nh (Bank)                | 1              | GP+wrapper [56]                                 |
|                | CPMP-2015-runtime-regression   | 1              | GP+wrapper [56]                                 |
|                | (CPMP)                         |                |                                                |
|                | Fri_c0_100_25 (Fri)            | 1              | GP+wrapper [56]                                 |
|                | MIP                            | 1              | GP+wrapper [56]                                 |
|                | Mtp                            | 1              | GP+wrapper [56]                                 |
|                | Selwood                        | 1              | GP+wrapper [56]                                 |
|                | Debutanize                     | 1              | GP+tree vector [52]                             |
|                | Weather_Izmir                  | 1              | GP+tree vector [52]                             |
|                | Kin8nm                         | 2              | GP+wrapper [55], GP+tree vector [52]            |
|                | Pol                            | 1              | GP+wrapper [55]                                 |
|                | Quake                          | 1              | GP+wrapper [55]                                 |
| UCI            | Airfoil-self-noise (Airfoil)   | 1              | GP+wrapper [55]                                 |
|                | Arrhythmia                     | 1              | DE+clustering [48]                              |
|                | Audiology                      | 1              | GA [17]                                         |
|                | Australian                     | 2              | MOGA-II [21], GA [18]                           |
|                | Auto mpg                       | 3              | PSO+covariance matrix [32],                    |
|                |                                |                | PSO+ECM+AAELM [44], GP+wrapper [55], GP+tree vector [52] |
|                | Automobile                     | 1              | DE+clustering [48]                              |
|                | Autos                          | 1              | GA [17]                                         |
|                | Balance scale                  | 1              | ABC+bayesian [25]                               |
|                | Body fat                       | 2              | PSO+covariance matrix [32],                    |
|                |                                |                | PSO+ECM+AAELM [44]                              |
|                | Boston housing                 | 2              | PSO+covariance matrix [32],                    |
|                |                                |                | PSO+ECM+AAELM [44]                              |
|                | Breast cancer                  | 3              | FCM+PSO [39], [42]                              |
|                |                                |                | KNN+LAHCAWOA [31]                               |
|                | Breast tissue                  | 1              | KNN+LAHCAWOA [31]                               |
|                | Bupa                           | 1              | FCM+PSO [39]                                    |
|                | Car                            | 1              | ABC+bayesian [25]                               |
|                | Census-Income (KDD)            | 1              | ABC+bayesian [24]                               |
|                | CCN                            | 2              | GP+wrapper [55], GP+tree vector [52]            |
|                | Cleveland heart disease        | 2              | GA [17], FCM+PSO [42]                           |
|                | Colonoscopy                    | 1              | KNN+LAHCAWOA [31]                               |
|                | Concrete                       | 2              | GP+wrapper [55], GP+tree vector [52]            |
| Dataset                  | Best approaches                                      |
|-------------------------|-----------------------------------------------------|
| Contraceptive           | MOGA-II [21]                                         |
| Covertype               | ACO+bayesian [24]                                    |
| Credit approval         | DE+clustering [48]                                   |
| CSM                     | GFM+PSO [43]                                         |
| Ecoli                   | MOGA-II [21]                                         |
| ENB2012                 | GP+wrapper [55], GP+tree vector [52]                 |
| Forest fires            | PSO+covariance matrix [32], PSO-ECM+AAELM [44], GA+MLP, SA+MLP, PSO+MLP, RF+MLP [58], GP+wrapper [55], GP+tree vector [52] |
| German                  | MOGA-II [21], GA [18]                               |
| Glass                   | FCM+SVR+GA [37], MOGA-II [21]                        |
| Haberman                | FCM+SVR+GA [37]                                      |
| Health records          | GA+MLP, SA+MLP, PSO+MLP, RF+MLP [58]                 |
| Heart                   | GA [18]                                             |
| Heart disease           | DE+clustering [48], GA [18]                         |
| Hepatitis               | GA [17], DE+clustering [48], GA+SOM [38]           |
| Horse colic             | DE+clustering [48], GA+SOM [38]                     |
| Housevotes              | DE+clustering [48]                                   |
| Hsv                     | GFM+PSO [43]                                         |
| Imports-85              | GP+wrapper [55], GP+tree vector [52]                 |
| Individual household electric power consumption | ACO+bayesian [24], Max–min ACO + bayesian [26], [27] |
| Insurance company benchmark | MAIS+GA [50]                                      |
| Ionosphere              | GA [18]                                             |
| Iris                    | FCM+GA [35], PSO+covariance matrix [32], FCM+SVR+GA [37], MOGA-II [21], PSO-ECM+AAELM [44], MOGA-II [22] FCM+PSO [39], [42] |
| KDD Cup 1998 Data       | Max–min ACO+bayesian [27]                            |
| Libras movement         | GP+wrapper [55]                                      |
| Liver                   | KNN+LAHCAWOA [31]                                    |
| Liver-disorder          | GP+wrapper [55]                                      |
| Localizations data for person activity | ACO+bayesian [24], Max–min ACO+bayesian [26] |
| Lung-cancer             | GA [17]                                             |
| Lymph                   | MOGA-II [21]                                         |
| Magic                   | MOGA-II [21]                                         |
| Mammographic masses     | GA [17], DE+clustering [48], GA+ARO [51]           |
| Marketing               | DE+clustering [48], GFM+PSO [43]                    |
| Musk1                   | FCM+SVR+GA [37]                                      |
| New-thyroid             | FCM+GA [35], MOGA-II [21]                           |
| Nursery                 | ABC+bayesian [25]                                    |
| Ozone                   | GP+wrapper [55], GP+tree vector [52]                 |
| Parkinson’s disease     | KNN+LAHCAWOA [31]                                    |
| Pima Indian **          | MOGA-II [21], GA [18], PSO+covariance matrix [32], PSO-ECM+AAELM [44], GA+ARO [51] |
| Poker hand              | ACO+bayesian [24], Max–min ACO + bayesian [26]      |
| Saheart                 | GA [18]                                             |
| Satimage                | MOGA-II [21]                                         |
| SECOM                   | GMSA+MPSO+WKNN [30]                                  |
| Shuttle                 | MOGA-II [21]                                         |
| SkillCraft1             | GP+wrapper [55], GP+tree vector [52]                 |
| Skin segmentation datasets | Max–min ACO+bayesian [26]                           |
| Sonar                   | GA [18], MOGA-II [22], GA+SOM [38]                  |
| Spanish                 | PSO+covariance matrix [32], PSO-ECM+AAELM [44]      |
| Spectf heart            | PSO+covariance matrix [32], PSO-ECM+AAELM [44], GA [18] |
| Temperature             | GFM+PSO [43]                                         |
| Thoraric                | KNN+LAHCAWOA [31]                                    |
| Tic-tac-toe             | MOGA-II [21]                                         |
| Turkish                 | PSO+covariance matrix [32], PSO-ECM+AAELM [44]      |
| UK bankruptcy           | PSO+covariance matrix [32], PSO-ECM+AAELM [44]      |
| UK credit               | PSO+covariance matrix [32], PSO-ECM+AAELM [44]      |
| Unseen credit           | GA+MLP, SA+MLP, PSO+MLP, RF+MLP [58]                 |
| US census data (1990)   | Max–min ACO+bayesian [26]                            |
| Dataset Description                                      | Model                                         | Note                                                                 |
|-----------------------------------------------------------|-----------------------------------------------|----------------------------------------------------------------------|
| Vertebral_column                                          | MOGA-II [21]                                  |                                                                      |
| Wdbc                                                      | GA [18], GA+SOM [38]                          |                                                                      |
| Website phishing                                          | ABC+bayesian [25]                             |                                                                      |
| Wine                                                      | MOGA-II [21], FCM+GA [35], PSO+covariance matrix [32], FCM+SVR+GA [37], PSO-ECM+AAELM [44] |                                                                      |
| Yacht hydrodynamics                                       | GP+wrapper [55], GP+tree vector [52]          |                                                                      |
| Yeast                                                     | FCM+SVR+GA [37]                               |                                                                      |
| Zoo                                                       | MOGA-II [22]                                  |                                                                      |
| Chu et al. [64] Microarray (Spo)                          | GA+KNN [29]                                   |                                                                      |
| Clare and King [65] Seq                                   | GA+KNN [29]                                   |                                                                      |
| ECBDL competition ECDL14 (ROS)                            | Max–min ACO+bayesian [27]                     |                                                                      |
| Gasch et al. [66] Microarray (Gasch2)                     | GA+KNN [29]                                   |                                                                      |
| Geo website                                               | Acute Myeloid Leukemia (AML)                  | PSO [20]                                                            |
| Germany                                                  | Forest Etteningen                             | RF+GA [49]                                                          |
| Harbin, China    Hourly traffic volume                    | FCM+GA [36]                                  |                                                                      |
| Harvard University Gene expression (Yeast)                | ACO+clustering [34]                           |                                                                      |
| IUUM Medical Centre Creatinine                           | PSO+levy flight [53]                          |                                                                      |
| Jahan Daneshgahi Research Center                          | Adult T-cell leukemia/lymphoma Gastric cancer | MOPSO [23]                                                           |
| Malaysian Meteorological Department                       | Malaysia meteorological                       | SC-FITNET [59]                                                      |
| Melbourne, Australia Yarra river basin                    | GP [19]                                       |                                                                      |
| Meteoblue website Basel weather                           | SC-FDO+MLP [60]                               |                                                                      |
| Minqin County, China Monthly groundwater level            | SOM+FOA+LSSVM [47]                           |                                                                      |
| Omid Hospital, Iran Breast cancer                        | Bayesian+tensor+chaotic PSO [28]              |                                                                      |
| Pascal Large Scale Learning Challenge                     | Epsilon                                      | Max–min ACO+bayesian [27]                                            |
| Princeton University Medical expenditure panel survey     | MAIS+GA [50]                                  |                                                                      |
| PKDD discovery challenge Hepatitis patient Thrombosis patient | IDW+TR+PSO [33]                              |                                                                      |
| South Africa South African insurance (SAI)                | MAIS+GA [50]                                  |                                                                      |
| Texas Texas insurance                                     | MAIS+GA [50]                                  |                                                                      |
| Thailand Thai dengue                                      | PSO+K-means+ontology model [46]               |                                                                      |
| University in Bogotá-Colombia Student information         | GA [16]                                       |                                                                      |
| China Byproduct Gas Flow Byproduct gas flow                | ELM+PSO+FCM [45]                              |                                                                      |
| China Nuclear Power Plant Hourly radiation dose rate      | PSO+LSSVM [54]                                |                                                                      |
| Xujiahui, China Hourly traffic                           | GSO+MLP [57]                                  |                                                                      |

Note: ** Pima Indian dataset is no longer available due to permission restrictions.

Figure 4 further shows the distribution of studies according to the number of the used dataset. As illustrated in Figure 4, nearly 41.7% of the articles used a minimum of one dataset, while others utilized multiple datasets. The number of datasets used in comparing algorithms varied from one to 15 datasets.

FIGURE 4. Distribution of studies based on number of used dataset.
2) MISSING MECHANISMS

From the findings, missingness can be grouped into two categories: real missing and synthetic missing datasets. A real missing dataset has the original missing data values, in which it does not include any synthetic or artificial missing ratios in the dataset. A synthetic missing dataset contains artificial missing ratios that have been inserted into the dataset according to the missing mechanisms. Nearly 79.2% (38/48) of studies in the last decade used synthetic datasets to evaluate imputation performance, while only 8.3% (4/48) used real missing datasets, and 6 studies did not clarify the missingness.

Among the synthetic missing datasets, MAR missing mechanism is the most popular mechanism, with 13 studies accounting for 27.1% (13/48) of the studies, followed by MCAR (20.8%, 10/48), MCAR+MAR (12.5%, 6/48), MCAR+MAR+MNAR (10.4%, 5/48). At the same time, the least attention is paid to MNAR missing mechanism. Dealing with the MNAR mechanism is complex and challenging [67]. The distribution of studies based on the missing mechanism is depicted in Figure 5.

3) MISSING RATES

The missing rates used in the experiment can be divided into three categories: missing rates <= 30%, missing rates under 30% – 50%, and missing rates > 50%. Figure 6 shows the distribution of studies based on the missing rates. According to the findings, dataset with missing rates <= 30% category is the most frequently used missing rates for experimentation in the studies (45.8%), followed by 25% of the studies designed to impute missing rates under 30% – 50% category. However, nearly 14.6% of the studies did not reveal their missing rates for the experimentation. The works in [24]-[25], for example, used the Framingham heart dataset with real missing values, but the authors did not disclose the proportion of missing values in the dataset.
| Year | Studies | Techniques | Dataset | Instance | Missing dataset | Missing rates (%) | Metric | Selected Top Results |
|------|---------|------------|---------|----------|----------------|------------------|--------|----------------------|
| 2011 | [34]    | ACO + clustering | Gene expression (Yeast) | 2882 | Synthetic missing: MCAR, MAR, MNAR | 2, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90 | RMSE | *MCAR≈25-120; *MAR≈ 23-27 |
| 2015 | [17]    | GA         | Audiology, autos, Cleveland heart disease, hepatitis, lung-cancer, mammographic masses | 32 - 961 | Real missing | 1.98 - 98.23 | Accuracy | Hepatitis dataset, classifier C4.5: 91.42% |
| 2015 | [21]    | MOGA-II    | Australian, ecoli, german, iris, magic, new-thyroid, pima, satimage, shuttle, wine, contraceptive, glass, lymph, tic-tac-toe, vertebral_column | 148 - 6435 | Synthetic missing: MCAR, MAR | 5 - 87 | Accuracy | MOGAImpACC≈ 87% |
| 2017 | [47]    | SOM + FOA+ LSSVM | Monthly groundwater level | 51 | Synthetic missing: MCAR | 10, 20, 30, 40, 50, 60, 70, 80 | CV-MAPE | Average: 5.6 |
| 2018 | [48]    | DE + clustering | Arrhythmia, automobile, credit approval, heart disease, hepatitis, horse-colic, housevotes, mammographic, marketing, ozone | 155 - 8993 | Real missing | 5 - 100 | Accuracy | Ozone dataset, KnnFFsCI imputation: 97.03% |
| 2021 | [52]    | GP + tree vector | Yacht, forest, EMB2012, concrete, weather_Izmir, debutanizer, kin8nm | 308 - 8191 | Synthetic missing: MAR | 30 | Relative square error (RSE) | Weather_Izmir dataset: 0.0312; Imports-85 dataset: 0.3175 |
| 2021 | [60]    | SC-FDO + MLP | Basel weather | 13057 | Synthetic missing: MCAR | 10, 20, 30, 40, 50, 60, 70, 80, 90 | R | Average R: 90% |
Nevertheless, the missing rates greater than 50% category received the least attention, accounting for 14.6% (7/48) of the studies. The detailed metaheuristic techniques for dealing with high missing rates are presented in TABLE 6. The techniques include ACO clustering for imputing gene expression database [34], GA imputation for infilling missing multi-attribute dataset [17], MOGA-II proposal for estimating missing data pattern in classification [21], data imputation of spatio-temporal underground water [47], DE clustering and feature selection with incomplete data [48], GP+tree vector imputer model for instance selection and symbolic regression on incomplete data [52] and SC-FDO based MLP trainer for missing rainfall time series imputation [60]. Among these imputation techniques, 6 out of 7 studies employed small-scale datasets of less than 10,000 instances. In general, the proposed approaches produced comparable results for the MVI tasks. Moreover, most high missing rate studies investigated MVI, assuming that the missing data mechanism is MCAR and MAR.

On the other hand, the work [60] utilized a large-scale dataset (over 10,000 instances) to fill in gaps for missing rainfall data. To sum up, the MVI studies need to be addressed from the three aspects, as illustrated in Figure 7. Therefore, future researchers need to identify the study's dataset characteristics, missing mechanisms, and missing rates.

C. EVALUATION METRICS

To answer RQ3, this subsection identifies the most often used metrics for evaluating the MVI's performances.

As illustrated in Figure 8, the nine most frequently used metrics for evaluating the performance of MVI were identified as the root mean square error (RMSE), accuracy,
correlation coefficient (R), mean square error (MSE), mean absolute error (MAE), error, mean absolute percentage error (MAPE), relative accuracy (RA), and specificity. Furthermore, 70.3% of the selected studies used these metrics. Many of the metrics are rarely used by the authors; therefore, these metrics have been categorized as ‘Others’.

The RMSE is the most frequently used metric for evaluating imputation performance, mainly to measure the differences between the predicted variables and the actual variables. For example, the works in [19], [20], [33], [36], [45], [46], [59], [60], to name a few, implemented this metric to determine how concentrated the predicted time series variables would be around the line of the actual variables. This metric is widely reported in time series imputation literature, such as missing rainfall, groundwater level, traffic volume, byproduct gas flow, and radiation dose rate data. Nagarajan and Dhinesh Babu [31] also used this metric to measure the performance of imputation related to missing health datasets.

Other than that, accuracy is used to measure the performance of the imputation method with respect to classifier accuracy [17], [23], [27], [28], [40], [41], [48], [50]. The MOGA-II imputer [21] achieved an accuracy of 82.98%, outperforming the GA imputer [18] and GA+ARO imputer [51] when handling missing values for the pima Indian dataset. Moreover, using the naïve Bayes classifier, the GA+ARO imputer [51] achieved the highest accuracy of 85% compared to GA imputer at 83.07% [17], and DE clustering imputer at 80.82% [48] for missing mammographic masses dataset. Another standard metric is error for summarizing the performance of imputation and classification models. For example, researchers adopted error metric to measure the classification errors of the proposed imputation models in the missing iris dataset [22], [35], [39], poker hand dataset [24], [26], website phishing dataset [25] and health datasets [31]. On the other hand, RA is an indicator of how many estimations fail within a standard range [36], [37], [45], [61], [62], while specificity (also true negative rate) refers to the proportion of sample without the condition but obtained a negative result [18], [23], [28], [43].

The R metric assesses the linear correlation between predicted and actual values. A higher R-value implies a better imputation performance. The works of [19], [36], [54], [58]–[60] used R to assess the correlation and association of the predicted and actual values for infilling missing values in the river basin, weather, traffic volume, and forest fire datasets. MSE is another metric for assessing the mean squared difference between predicted and actual values. For example, Garg et al. [63] measured their proposed DL-GSA imputation with the works in [61] and [62] in terms of R and MSE. The results revealed that the DL-GSA imputation method produced more substantial correlation results and lower MSE than the works in [61] and [62]. Some researchers also adopted MAE to measure the proposed imputation methods in terms of the average magnitude of the errors for continuous variables [53], [58]–[60].

Nevertheless, the shortcoming of the MAE metric is that it does not consider the direction of the mean error. As opposed to this shortcoming, Willmott [68] suggested that comparing average model performance error should use MAE because MAE is a natural measure of average error magnitude. In some instances, MAPE is essential to assess the prediction accuracy of the imputation models. Zhang [57] used MAPE to evaluate imputation results in missing Spatio-temporal data. Concerning MAPE, the PSO-ECM+AAELM imputer [44] outperformed the PSO+covariance matrix imputer [32] for all 12 datasets, such as autompg, body fat, boston housing, forest fires, iris, pima Indian, Spanish, spectf, Turkish, UK bankruptcy, UK credit and wine datasets.

IV. DISCUSSION

This section discusses the research trends and potential opportunities in the metaheuristic approach for handling and optimizing MVI.

A. THE MVI APPROACHES

In reference to the RQs, which attempt to identify the existing metaheuristic techniques used for handling and optimizing MVI, it can be revealed that most techniques used to handle missing values were hybrid metaheuristics with clustering or neural networks. Each of the hybrids has characteristics that make it a good fit for a particular problem. For example, the hybrids of deep-autoencoder and metaheuristics provide good results in imputing high-dimensional handwritten digits. In particular, the DL-GSA [63] imputer model was faster and more accurate than the DL-CS [61] and DL-BAT [62]. However, the computational times of the hybrids MLP and metaheuristics (GA+MLP, SA+MLP, and PSO+MLP) [58] were relatively shorter than the DL-GSA, DL-BAT, and DL-CS approaches. On the other hand, the work in [59] indicated that the hybrid function of fitting neural network and metaheuristics (SCFITNET) yielded more accurate estimates than LSTM imputer model for missing rainfall data when correlation coefficient, mean absolute error, and root mean square errors were taken into account. Therefore, selecting the suitable imputer model best suited for the incomplete datasets is essential. Additionally, the hybridization of the state-of-art metaheuristic and neural networks could be of interest for the researchers, therefore providing new studies.
Typically, researchers perform a series of studies to fine-tune parameters in imputer models, which requires considerable effort. For instance, metaheuristic parameters [24], [26], such as the population size and the iteration count, require fine-tuning; parameters in neural network models [61]–[63] are the number of hidden layers in the neural network and the number of neurons in the hidden layer; and parameters in clustering [39] such as the fuzzification parameter, the number of clusters, and the number of nearest neighbors all require fine-tuning.

Consequently, several studies [69]–[71] investigate automatic parameter tuning methods to optimize the algorithm's performance. However, there is no universally accepted guideline for selecting the optimal set of parameters to achieve the best performance. Therefore, future research could consider a semi-automatic or automatic parameter tuning approach for a given context and domain.

Dataset scales (the number of instances) in a dataset influence the imputation performance. Data resources with few instances may cause imputed values to be underestimated or overestimated. Therefore, researchers must expand the size of the databases, as small-scale datasets can lead to biases and a lack of generalization. Furthermore, training on a large-scale and high-dimensional dataset is difficult due to computational complexity. For example, neural networks, deep learning algorithms need many data to improve accuracy. Hence, dimensionality reduction approaches can help reduce computational costs and improve the accuracy of imputation performance.

On the other hand, imputation models [52] built on a relatively small number of instances (< 300) or a large number of instances (> 8191) were ineffective and inaccurate. For this reason, researchers need to comprehend the requirements in both the problem and solution domains before proposing an imputer model.

Furthermore, less attention has been paid to real-world datasets from industries or agencies. Therefore, real-world datasets from industries or agencies with larger scales (over 10,000 instances) and higher dimensions might be the areas worth exploring by future researchers.

### D. THE MISSING MECHANISMS

The approaches of handling incomplete data are associated with the missing mechanisms. MAR and MCAR are the two most frequently used for evaluating imputation performance among the missing mechanisms. However, the MNAR missing mechanism receives the least attention.

Domain-based imputation approaches are developed to deal with the problem of incomplete data. It is not envisaged that some features are missing for all patients in medical datasets. In contrast, large scale of datasets (over 10,000 instances) were used in the works of [26], [27] (discrete, continuous data type), and [59], [60] (continuous data type). Meanwhile, handwriting datasets with high dimensions and scales [61]–[63] were used.

Dataset scales (the number of instances) in a dataset influence the imputation performance. Data resources with few instances may cause imputed values to be underestimated or overestimated. Therefore, researchers must expand the size of the databases, as small-scale datasets can lead to biases and a lack of generalization. Furthermore, training on a large-scale and high-dimensional dataset is difficult due to computational complexity. For example, neural networks, deep learning algorithms need many data to improve accuracy. Hence, dimensionality reduction approaches can help reduce computational costs and improve the accuracy of imputation performance.
possibility of missing for all days when hardware failure occurs at a specific gauging station. In this case, the MAR and MNAR missing mechanisms are appropriate for evaluating imputation performance on incomplete medical datasets.

The missing rainfall feature of one gauging station does not influence the other gauging stations. Therefore, a domain-based imputation approach and missing mechanism for a given context should be investigated further to improve the adaptability and accuracy of the imputation models. For this reason, the MCAR missing mechanism is appropriate for evaluating the incomplete rainfall datasets.

E. THE MISSING RATES
The ability of imputer approaches to handle complexity is tested using different percentages of missing values. Most studies reported that at lower missingness, the performances of MVI are relatively better. Imputation errors increased when missing rates increased, for examples in [22], [24], [25], [28], [33], [34], [45], [46]. In addition, the percentages of missingness greatly influenced the work in [26], [35].

The findings also indicated that synthetic datasets with missing rates less than 30% are the most frequently used missing rates for experimentation in studies (45.8%), while only 14.6% of the studies considered missing rates greater than 50%. However, the missing rates could be larger than 50% in real-world problems. Therefore, this SLR suggests designing MVI methods that can deal with low and high missingness problems, for example, missing rates of 10% - 90%. These findings also agree with other work [10] that imputation studies with more significant missing rates would be more practical.

F. THREATS TO VALIDITY
Four potential threats to validity should be considered to support the findings of this SLR: construct, internal, external, and conclusion validity. To achieve maximum construct validity, we conducted this literature review following Kitchenham's guidelines [13] and performed analyses in response to research questions, quality assessment, and inclusion and exclusion criteria. However, the relevance of various terms associated with the missing could constrain our findings. We attempt to maximize internal validity by applying all missing terms associated with imputation techniques and datasets as described in Table 4 and Table 5. Additionally, we seek to maximize internal validity by employing an exhaustive manual and automated search strategy to ensure the paper selection was unbiased. Further, external validity considers whether our findings can be generalized to other studies. In this study, we emphasize MVI designs and methods of metaheuristic techniques exclusively, holding the other paradigms for future research.

Finally, data extraction was carried out to ensure the conclusion's validity by adhering to the review protocols, including the research questions, quality assessment, inclusion criteria, search strategy, and study selection [15]. Other review protocols could be used, which may increase or decrease research bias and lead to different findings.

V. CHALLENGES IN IMPLEMENTING MISSING VALUE IMPUTATION DESIGNS AND METHODS
There will be challenges with any new research method, especially in identifying the appropriate approaches for a wide range of research questions and experimental designs. Careful planning and consideration are required to reduce the impact of missing values and improve data quality. The following section discusses some roadblocks to implementing MVI and tentative guidelines.

A. IMPUTATION PERFORMANCES AND COMPUTATIONAL COST
One of the significant MVI challenges is the expensive computational time, especially with large-scale and high-dimensional datasets. Data normalization, feature selection, or feature extraction can be employed to reduce the computational cost. For example, [48] demonstrated that feature selection significantly reduced the computational time of imputation while improving the imputation and classification accuracy.

B. UNPLANNED MISSING VALUE
In some studies, data with missing values were removed [73]–[75]. The works in [3], [76], [77] also removed missing time-series data from experiments. However, it is important to note how the authors dealt with the records’ continuity because accurate forecasting relies on continuous time-series records. Furthermore, Hussain et al. [78] reported that many missing data entries made it challenging to accurately impute the electric power consumption data. Only 60.11% of the total consumers with null entries lower than 200 were considered for MVI, whereas 39.89% of the customer records were removed from the experiment. However, removing missing values from observations results in a reduction in sample representativeness. The effects of unintentional missing values can induce biases in parameter estimates and uncertainty, which can be mitigated by adopting an effective MVI procedure and design plan.

C. OPTIMAL MISSING VALUE IMPUTATION APPROACHES
This study also revealed no definitive answer on which method is the best to date for all the missingness. The adoption of MVI approaches depends on many factors: data characteristics, missingness mechanisms, the proportion of missing values, dependent and independent variables, dataset volume, computational time, and domain applications.
Consequently, the existing reports of MVI studies are of great worth assisting future researchers in developing an effective MVI strategy. However, 14.6% of the studies did not report on missing rates, whereas 20.8% of the studies (10/48) did not clarify the used missingness mechanism. This information is a valuable factor when planning for the experimental design of MVI. Therefore, an overview of the recommended guidelines in addressing, managing, and reporting MVI studies is outlined in Figure 9.

The MVI strategic planning process begins with the collection of incomplete datasets. It is crucial to identify the three main aspects of incomplete datasets: dataset characteristics, missing mechanisms, and missing rates. The next step is the selection of MVI approach. Having a clear justification of the chosen strategy, the potential impact of imputation, and computational cost are crucial to the success of MVI method. Without a clear direction, the MVI strategy may stall or even fail. Data normalization, feature selection or feature extraction method could be considered to improve the performances of the MVI approach.

Researchers can then use complete or incomplete training datasets to construct optimal imputer models. The incomplete dataset can be real missing or synthetic missing dataset. Training and testing dataset design, variables with missing data, missing rates, missing mechanism, and dataset characteristics should be thoroughly reported. Researchers should train the imputer models on one dataset and test them on another dataset to verify the robustness of the proposed imputer models. A set of performance metrics is used to measure the effectiveness and efficiency of the MVI method. The commonly used metrics are RMSE, accuracy, R, MSE, and MAE. Statistical analysis such as Wilcoxon signed-rank test [21], [32], Wilcoxon rank-sum test [37], and Friedman test [21] can be performed to assess the significance of the proposed MVI approach. Finally, we suggest that the researchers to report the three factors affecting MVI in detail (dataset characteristics, missing mechanisms, and missing rates), training and testing procedures, measurement metrics, and the findings of the studies.

Additionally, the reporting could couple with the discussion of the impact and challenges of the MVI, which will increase the overall confidence in the study.

The planned MVI procedures and strategies can raise statistical power and model convergence compared to employing a complete case analysis [79]. Preparing for missing values before starting an experiment can also help avoid the problems of nonrandom missing data, leading to significant bias and invalid statistical inferences [2], [80]. Furthermore, researchers can use the planned MVI design in conjunction with missing data procedures to increase the quality and scope of the study and lower research costs. Researchers might minimize the study cost by strategically implementing an effective MVI design.

FIGURE 9. A guide to address, manage and report missing value imputation studies.

VI. CONCLUSION

In recent years, MVI for incomplete datasets has grown in popularity to improve data quality, statistical power and reduce bias in data science applications. In this study, we conducted a SLR to examine the existing metaheuristic techniques used for handling and optimizing missing value imputation over the last ten years. This SLR is also
concerned with establishing guidelines for researchers in the domain to understand MVI technologies and designs better. This study concentrated on three major scientific databases: IEEEExplore, ScienceDirect, and Scopus. The findings of this SLR revealed that the hybridizations of metaheuristics with clustering or neural networks are the most used MVI approaches. The review indicates that the hybrid metaheuristic is a promising field of study for solving various imputation problems.

Additionally, we discovered that the synthetic missing dataset is the most frequently used incomplete dataset for evaluation, and RMSE is the most used metric for evaluating the performance of the proposed MVI. When handling missing data, the three aspects to consider are the dataset characteristics, missing mechanisms, and missing rates. This review also addresses MVI perspectives, challenges, and opportunities. An optimal imputer approach by domain-based approaches should be investigated further. However, designing a planned MVI design and method to expand the quality of study scope remains a significant challenge. Therefore, the literature provides an overview of recommended guides for planning MVI designs and methods, which serve as a toolkit for developing an effective MVI strategy.

**APPENDIX**

| Acronym  | Full form                                                      |
|----------|----------------------------------------------------------------|
| AAELM    | Autoassociative extreme learning machine                      |
| ABC      | Artificial bee colony                                        |
| ACO      | Ant colony optimization                                       |
| ARO      | Asexual reproduction optimization                              |
| BAT      | Bat algorithm                                                 |
| CS       | Cuckoo search                                                 |
| DE       | Differential evolution                                        |
| DL       | Deep learning                                                 |
| ECM      | Evolving clustering method                                     |
| ELM      | Extreme learning machine                                      |
| FA       | Firefly algorithm                                             |
| FCM      | Fuzzy C-means                                                 |
| FDO      | Fitness dependent optimizer                                   |
| FOA      | Fruit fly optimization algorithm                              |
| GA       | Genetic algorithm                                             |
| GFM      | Granular fuzzy models                                         |
| GMMSA    | Gaussian mutation simulated annealing                          |
| GP       | Genetic programming                                           |
| GSA      | Gravitational search algorithm                                 |
| GSO      | Group search optimization                                     |
| IDW      | Inverse distance weight                                       |
| KNN      | K-nearest neighbor                                            |
| LAHCAWOA| Late acceptance hill climbing algorithm + whale optimization algorithm |
| LSSVM    | Least squares support vector machine                           |
| LSTM     | Long short-time memory                                        |
| MAE      | Mean absolute error                                           |
| MAPE     | Mean absolute percentage error                                 |
| MAR      | Missing at random                                             |
| MCAR     | Missing completely at random                                  |
| MICE     | Multivariate imputation by chained equations                  |
| MAIS     | Multi-layered artificial immune system                         |
| MLP      | Multilayer perceptron                                         |
| MNAR     | Missing not at random                                         |
| MOGA-II  | Multi objective genetic algorithm-II                          |
| MOPSO    | Multi objective particle swarm optimization                    |
| MPSO     | Memetic particle swarm optimization                            |
| MSE      | Mean square error                                             |
| MVI      | Missing value imputation                                       |
| NSGA-II  | Non-dominated sorting genetic algorithm-II                    |
| PSO      | Particle swarm optimization                                    |
| R        | Correlation coefficient                                       |
| RA       | Relative accuracy                                             |
| RF       | Random forest                                                 |
| RMSE     | Root mean square error                                        |
| RQ       | Research question                                             |
| SA       | Simulated annealing                                           |
| SCA      | Sine cosine algorithm                                          |
| SC-FDO   | Sine cosine-fit dependent optimizer                            |
| SC-FITNET| Sine cosine function fitting neural network                    |
| SLR      | Systematic literature review                                   |
| SOM      | Self-organizing map                                           |
| SVR      | Support vector regression                                     |
| TR       | Tolerance rough set                                            |
| WKNN     | Weighted K-nearest neighbor                                    |
| WOA      | Whale optimization algorithm                                   |

**REFERENCES**

[1] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0206-3.

[2] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, 3rd ed. Wiley, 2014.

[3] K. J. Bormann, S. Westra, J. P. Evans, and M. F. McCabe, “Spatial and temporal variability in seasonal snow density,” *J. Hydrol.*, vol. 494, pp. 63–73, 2013, doi: 10.1016/j.jhydrol.2013.01.032.

[4] A. Tatar, I. Askarova, A. Shafiei, and M. Rayhani, “Data-Driven Connectionist Models for Performance Prediction of Low Salinity Waterflooding in Sandstone Reservoirs,” *ACS Omega*, vol. 6, no. 47, pp. 32304–32326, 2021, doi: 10.1021/acsomega.1c05493.

[5] B. N. I. Eskelson, T. M. Barrett, and H. Temesgen, “Imputing mean annual change to estimate current forest attributes,” *Silva Fenn.*, vol. 43, no. 4, pp. 649–658, 2009, doi: 10.14214/sf.185.

[6] H. Turabieh, M. Mafarja, and S. Mirjalili, “Dynamic Adaptive Network-Based Fuzzy Inference System (D-ANFIS) for the Imputation of Missing Data for Internet of Medical Things Applications,” *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9316–9325, 2019, doi: 10.1109/JIOT.2019.2926321.

[7] P. C. Chiu, A. Selamat, O. Krejcar, and K. K. Kuok, “Missing rainfall data estimation using artificial neural network and nearest neighbor imputation,” *Front. Artif. Intell. Appl.*, vol. 318, pp. 132–143, 2019, doi: 10.3233/FIA190044.

[8] P. C. Chiu, A. Selamat, and O. Krejcar, “Infilling missing rainfall and runoff data for Sarawak, Malaysia using gaussian mixture model based K-nearest neighbor imputation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11606 LNAI, pp. 27–38, 2019, doi: 10.1007/978-3-030-22999-3_3.

[9] J. Tang, X. Zhang, T. Yu, and F. Liu, “Missing traffic data imputation considering approximate intervals: A hybrid structure integrating adaptive network-based inference and fuzzy rough set,” *Phys. A Stat. Mech. its Appl.*, vol. 573, p. 125776, 2021, doi: 10.1016/j.physa.2021.125776.

[10] W. C. Lin and C. F. Tsai, “Missing value imputation: a review and
analysis of the literature (2006–2017),”*Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, 2020, doi: 10.1007/s10462-019-09709-4.

[11] 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, 2021, doi: 10.1108/DATA-12-2020-0298.

[12] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, “Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021),” *Informatics Med. Unlocked*, vol. 27, p. 100799, 2021, doi: 10.1016/j.imu.2021.100799.

[13] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” *Tech. Rep. EBSE 2007-001 Keele Univ. Durham Univ. Jt. Rep.*, 2007, [Online]. Available elsiever.com/_data/promis_mscis/525444systematicreviewsguide.pdf.

[14] N. Genc-Nayebi and A. Abran, “A systematic literature review: Opinion mining studies from mobile app store user reviews,” *J. Syst. Softw.*, vol. 125, pp. 207–219, 2017, doi: 10.1016/j.jss.2016.11.027.

[15] L. Yang et al., “Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective,” *Inf. Softw. Technol.*, vol. 130, no. August 2020, p. 106397, 2021, doi: 10.1016/j.infsof.2020.106397.

[16] I. C. Figuerola Garcia, D. Kalenatic, and C. A. Lopez Bello, “Missing data imputation in multivariate data by evolutionary algorithms,” *Comput. Human Behav.*, vol. 27, no. 5, pp. 1468–1474, 2011, doi: 10.1016/j.chb.2010.06.026.

[17] F. M. F. Lobato, V. W. Tadaiesky, I. M. De Araújo, and A. L. Santana, “An Evolutionary Missing Data Imputation Method for Pattern Classification Categories and Subject Descriptors,” pp. 1013–1019.

[18] C. Panse, M. Kshirsagar, D. Raje, and H. Hiary, “Evolmputer: An evolutionary approach for Missing Data Imputation and feature selection in the context of supervised learning,” *Knowledge-Based Syst.*, vol. 236, p. 107734, 2022, doi: 10.1016/j.knosys.2021.107734.

[19] C. Sivapragasam, N. Muttill, M. C. Jeselio, and S. Visweswaran, “Infilling of Rainfall Information Using Genetic Programming,” *Aquat. Procedia*, vol. 4, no. 1cwcroce, pp. 1016–1022, 2015, doi: 10.1016/j.aqpro.2015.02.128.

[20] C. Parvez, H. A. Khorshidi, M. Kirley, and U. Aickelin, “Machine learning with max–min ant colony optimization,” *Informatics Med. Unlocked*, vol. 55, no. 4, pp. 558–585, 2021, doi: 10.1108/DTA-12-2020-0298.

[21] R. Sivaraj and R. Devi Priya, “Bayesian-based parallel ant system for missing value estimation in large databases,” *Trans. Inst. Meas. Control*, vol. 36, no. 8, pp. 1083–1092, 2014, doi: 10.1177/0142331214543291.

[22] 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. Inform.*, vol. 94, no. December 2018, p. 103190, 2019, doi: 10.1016/j.jbi.2019.103190.

[23] N. Krishna and V. Ravi, “Particle swarm optimization and covariance matrix based data imputation,” 2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013, vol. 2000507, 2013, doi: 10.1109/ICCIC.2013.6724232.

[24] J. Y. Nancy, N. H. Khanna, and K. Arpurthari, “Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework,” *Comput. Stat. Data Anal.*, vol. 112, pp. 63–79, 2017, doi: 10.1016/j.csda.2017.02.012.

[25] R. Veronez, F. O. De Franca, and F. J. Von Zaben, “Assessing the performance of a swarm-based biclustering technique for data imputation,” 2011 IEEE Cong. Evol. Comput. CEC 2011, pp. 386–393, 2011, doi: 10.1109/CEC.2011.5949644.

[26] D. Li, H. Gu, and L. Zhang, “A hybrid genetic algorithm-fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals,” *Soft Comput.*, vol. 17, no. 10, pp. 1787–1796, 2013, doi: 10.1007/s00500-013-0997-7.

[27] J. Tang, G. Zhang, Y. Wang, H. Fang, and F. Liu, “A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation,” *Transp. Res. Part C Emerg. Technol.*, vol. 51, pp. 29–40, 2015, doi: 10.1016/j.trc.2014.11.003.

[28] I. B. Aydilek and A. Arslan, “A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm,” *Inf. Sci. (Ny)*, vol. 233, pp. 25–35, 2013, doi: 10.1016/j.ins.2013.01.021.

[29] B. K. Khotimah and Y. D. Pramudita, “Initial center weight self organizing map using genetic algorithm,” *Proceeding - 6th Int. Technol. Int. Semin. ITIS 2020*, pp. 263–268, 2020, doi: 10.1016/9780075011820.9321044.

[30] L. Zhang, W. Lu, X. Liu, W. Pedrycz, C. Zhong, and L. Wang, “A Global Clustering Approach Using Hybrid Optimization for Improved Data Based Real Valued Recognition of Missing Value,” *Int. J. Intell. Syst.*, vol. 31, no. 4, pp. 297–313, 2016, doi: 10.1002/int.

[31] M. N. M. Salleh and N. A. Samat, “FCMPSO: An Imputation for Missing Data Features in Heart Disease Classification,” *IJOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, 2017, doi: 10.1088/1757-999X/226/1/012102.

[32] M. N. M. Salleh and N. A. Samat, “An Imputation for Missing Data Features Based on Fuzzy Swarm Approach in Heart Disease Classification,” in *International Conference on Swarm Intelligence*, 2017, pp. 285–292, doi: 10.1007/978-3-319-61833-3_30.

[33] N. A. Samat and M. N. M. Salleh, “A Study of Data Imputation Using Fuzzy C-Means with Particle Swarm Optimization,” in *International Conference on Soft Computing and Data Mining*, 2016, pp. 91–100, doi: 10.1007/978-3-319-51281-5_10.

[34] X. Hu, Y. Shen, W. Pedrycz, Y. Li, and G. Wu, “Granular Fuzzy Rule-Based Modeling With Fuzzy Complete Data Representation,” *IEEE Trans. Cybern.*, no. mi, pp. 1–14, 2021, doi: 10.1109/TCYB.2021.3071145.

[35] C. Gautam and V. Ravi, “Data imputation via evolutionary computation, clustering and a neural network,” *Neurocomputing*, vol. 156, pp. 134–142, 2015, doi: 10.1016/j.neucom.2014.12.073.

[36] X. Sun, Z. Wang, and J. Hu, “ELM-PSO-FCM based missing values imputation for byproduct gas flow data analysis,” *Proc. 2019 IEEE 3rd Int. Conf. Technol. Networking Electron. Autom. Control Conf. ITNEC 2019", no. Innee, pp. 56–59, 2019, doi: 10.1109/ITNEC.2019.8729038.

[37] N. Kamkhad, K. Jampachaisri, P. Siriyasatien, and K. Kesorn, “Toward semantic data imputation for a dengue dataset,” *Knowledge-Based Syst.*, vol. 196, p. 105803, 2020, doi: 10.1016/j.knosys.2020.105803.
PO CHAN CHIU is currently pursuing a Ph.D. degree in Computer Science from Universiti Teknologi Malaysia (UTM). She received an M.Sc. in information technology from the Universiti Malaysia Sarawak (UNIMAS), in 2010. She started her career as a Software Engineer for three years. She worked on several consultancy projects and developed software solutions to meet the needs of the woodworking industry. Currently, she is serving as a lecturer at UNIMAS. Her research interests include artificial intelligence, data analytics, optimization, and neural networks.

ALI SELamat is currently a Full Professor with Universiti Teknologi Malaysia (UTM), Malaysia. He has also been the Dean of the Malaysia Japan International Institute of Technology (MJITT), UTM, since 2018. An academic institution established under the cooperation of the Japanese International Cooperation Agency (JICA) and the Ministry of Education Malaysia (MOE) to provide the Japanese style of education in Malaysia. He is also a Professor with the Software Engineering Department, School of Computing, UTM and the Chair of the IEEE Computer Society Malaysia Section. He has published more than 120 research articles with IF JCR, with more than 2400 citations received in the Web of Science and h-index 26. His research interests include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, collective computational intelligence, strategic management, key performance indicator, and knowledge management. He is on the Editorial Board of the journal Knowledge-Based Systems (Elsevier).

ONDREJ KREJCAR is currently a Full Professor of systems engineering and informatics with the University of Hradec Králové, Czech Republic. He is also the Vice-Dean for science and research at the Faculty of Informatics and Management, UHK. He is also the Director of the Center for Basic and Applied Research, University of Hradec Králové. At the University of Hradec Králové, he is a guarantor of the Doctoral Study Programme in Applied Informatics, where he is focusing on lecturing on smart approaches to the development of information systems and applications in ubiquitous computing environments. His h-index is 20 (according to Web of Science), with more than 1500 citations received in the Web of Science. He has published more than 110 research articles with IF JCR. He has a number of collaborations throughout the world (e.g., Malaysia, Spain, U.K., Ireland, Ethiopia, Latvia, and Brazil). His research interests include control systems, smart sensors, ubiquitous computing, manufacturing, wireless technology, portable devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of interest is in biomedicine (image analysis), as well as biotelemetric system architecture (portable device architecture and wireless biosensors), and the development of applications for mobile devices with use of remote or embedded biomedical sensors.

Dr. Krejcar has also been a Management Committee Member substitute of the project COST CA16226, since 2017. In 2018, he was the 14th Top-Peer Reviewer in Multidisciplinary in the World according to Plumx. He is on the Editorial Board of Sensors (MDPI) with JCR Index and several other ESCI indexed journals. He has been the Vice-Leader and a Management Committee Member at WG4 of the project COST CA17136, since 2018. Since 2019, he has been the Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic, as a Regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019–2024). Since 2014, he has been the Deputy Chairman of the Panel 7 (Processing Industry, Robotics and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic.

KING KUOK KUOK is currently an Associate Professor at Swinburne University of Technology Sarawak Campus. He received his MEng from the UNIMAS in 2004 and Ph.D. from the UTM in 2010. He was the Field Engineer for Hydrological and Water Resources Branch, Department of Irrigation and Drainage, State of Sarawak, Malaysia from 2002 to 2009 and the Road, Civil and Structural Design Engineer at private companies for more than ten years. His research interests include water resources, water supply, information modeling.

SITI DIANAH ABDUL BUJANG received the B.S degree in Science (Computer Science) and M.S degree in Science from Universiti Teknologi Malaysia (UTM) in 2006 and 2010, respectively. She is currently pursuing the Ph.D. degree in Software Engineering at Malaysia-Japan International Institute of Technology, UTM in Kuala Lumpur. Her thesis focuses on the application of predictive analytics on student grade prediction in a higher education institution. From 2010 to 2019, she was the senior lecturer of Information and Communication Technology Department at Politecnic Sultan Idris Shah, Sabak Bernam, Selangor, Malaysia. She has experienced in developing the polytechnic curriculum for Diploma in Information Technology (Technology Digital) 2.5 years’ program. She is one of the book authors that contribute for the Department of Polytechnic and Community College Education. Her research interests include data analytics, predictive analytics, learning analytics, educational data mining and machine learning.

HAMIDO FUJITA (IEEE, Life Senior Member) is Emeritus Professor of Iwate Prefectural University, Takizawa, Japan. He is currently the Executive Chairman of i-SOMET Incorporated Association, Morioka, Japan. He is Highly Cited Researcher in Cross-Field for the year 2019 and 2020 in Computer Science field, respectively from Clarivate Analytics. He received Doctor Honoris Causa from Obuda University, Budapest, Hungary, in 2013 and received Doctor Honoris Causa from Timisoara Technical University, Timisoara, Romania, in 2018, and a title of Honorary Professor from Obuda University, in 2011. He is Distinguished Research Professor at the University of Granada, and Adjunct Professor with Stockholm University, Stockholm, Sweden; University of Technology Sydney, Ultimo, NSW, Australia; National Taiwan Ocean University, Keelung, Taiwan, and others. He has supervised Ph.D. students jointly with the University of Laval, Quebec City, QC, Canada; University of Technology Sydney; Oregon State University, Corvallis, OR, USA; University of Paris 1 Pantheon-Sorbonne, Paris, France; and University of Genoa, Italy. Dr. Fujita is the recipient of the Honorary Scholar Award from the University of Technology Sydney, in 2012. He has four international patents in software systems and several research projects with Japanese industry and partners. He is the Emeritus Editor-in-Chief for Knowledge-Based Systems, and currently Editor-in-Chief of Applied Intelligence (Springer). He headed a number of projects including intelligent HCI, a project related to mental cloning for healthcare systems as an intelligent user interface between human-users and computers, and SCOPE project on virtual doctor systems for medical applications. He collaborated with several research projects in Europe, and recently he is collaborating in Olimpia project supported by Tuscany region on Therapeutic monitoring of Parkinson disease. He has published more 400 highly cited Papers.
