A Systematic Literature Review On Missing Values : Research Trends, Datasets, Methods and Frameworks

Ismail Setiawan[1*], *Rahmat Gernowo*[2], and *Budi Warsito*[1]

[1] Department of Information System, Universitas Diponegoro, Indonesia
[1]Departement of Information Technology And System, Faculty Of Science And Technology, ‘Aisyiyah Surakarta University, Indonesia
[2]Department of Information System, Universitas Diponegoro, Indonesia
[3]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Indonesia

Abstract. Handling of missing values in data analysis is the focus of attention in various research fields. Imputation is one method that is commonly used to overcome this problem of missing data. This systematic literature review research aims to present a comprehensive summary of the relevant scientific literature that describes the use of the imputation method in overcoming missing values. The literature search method is carried out using various academic databases and reliable sources of information. Relevant keywords are used to find articles that match the research question. After selection and evaluation, 40 relevant articles were included in this study. The findings of this study reveal a variety of imputation approaches and methods used in various research fields, such as social sciences, medicine, economics, and others. Commonly used imputation methods include single imputation, multivariate imputation, and model-based imputation methods. In addition, several studies also describe a combination of imputation methods to deal with more complex situations. The advantage of the imputation method is that it allows researchers to maintain sample sizes and minimize bias in data analysis. However, the research results also show that the imputation method must be applied with caution, because inappropriate imputation decisions can lead to biased results and can affect the accuracy of the research conclusions. In order to increase the validity and reliability of research results, researchers are expected to transparently report the imputation method used and describe the considerations made in the imputation decision-making process. This systematic review of the literature review provides an in-depth view of the use of the imputation method in handling missing values. In the face of the challenge of missing data, an understanding of the various imputation methods and the context in which they are applied will be key to generating meaningful findings in various research fields.

[*] Corresponding author: author@email.org

1 Introduction

The handling of missing values is an important concern in research because the presence of missing data can cause bias in the analysis, reduce the validity of research results, and affect the accuracy of the conclusions drawn [1]–[4]. When missing data is not properly handled, it can lead to bias in statistical analysis. Research results based on incomplete data can lead to inaccurate conclusions and obscure the true relationship between variables. Missing value results in loss of valuable information from research samples [5], [6]. The more missing data, the more lost information, and this can reduce the efficiency and interpretation of research results [7]–[9]. Missing data can cause a decrease in the validity and reliability of research results. In some cases, missing data can disrupt the overall data structure and prevent proper analysis. When missing data only occur in certain groups or regions, research results may lose the ability to generalize to the population as a whole

There are many examples of cases or areas of research where missing value problems often arise [10]–[15]. In Health Research, medical and health data often contain missing values for several reasons, such as patients who miss several visits or do not complete all health questionnaires. In Survey and Social Studies research, in surveys and social studies, respondents may not answer some questions or leave some blank items in the questionnaire, resulting in missing values in the data [16], [17]. In economic research, in economic research, missing data can arise due to a number of factors, including errors in data collection or the absence of respondents from interviews [18], [19]. in Educational Research, in the field of education, missing data can occur when students do not fill in all the questions in a test or survey, or when school administration data is incomplete [20]. In Financial Analysis research, in financial analysis, missing data can occur in financial reports, such as company profit and loss or balance sheets, which can affect the evaluation of financial performance [21] [22].

The handling of missing values in research has a significant important impact, especially in producing accurate conclusions and better generalizations [23]. Following are some of the important impacts of handling missing values:

1. Accuracy of Statistical Analysis: When missing data is left without proper handling, the statistical analysis performed can be biased and lead to inaccurate conclusions. Statistical methods that use complete and valid data will provide more precise and reliable analysis results.
2. Better generalizations: In scientific research, the main goal is to make generalizations about the wider population based on the sample taken. If missing data is not handled properly, the generalization of research results may be unreliable, because the sample used may not reflect the population as a whole.
3. Validity and Reliability: Good handling of missing values will increase the validity and reliability of research results. Validity reflects the extent to which the measurement instrument actually measures the variable in question, while reliability reflects the consistency and accuracy of the measurement instrument. By handling missing values, data integrity can be maintained so that research results are more valid and reliable.
4. Research Efficiency: Handling missing values allows researchers to maximize the use of existing data. By filling in or estimating missing data, researchers can reduce information loss and increase the efficiency of data analysis.
5. Better Interpretation: Missing data can lead to erroneous analysis and inaccurate interpretation. By handling the missing values appropriately, researchers can be more confident in interpreting research results and provide better meaning to the findings found.

6. Reducing Research Bias: Missing values that are not addressed can lead to bias in the sample, and therefore, affect the research results. With proper handling, researchers can reduce bias that may arise and produce more objective findings.
7. Research Quality Improvement: By carefully handling missing values, research quality can be improved. Research that has complete and valid data will be methodologically stronger, thereby increasing confidence in the validity of research findings.

Overall, handling missing values in research is an important and critical step to ensure the reliability and validity of research results. By properly addressing the problem of missing data, researchers can produce more accurate analyses, stronger conclusions, and better generalizations to make meaningful contributions to their research fields.

2 Method

2.1 Literature Search Procedure

In order to carry out systematic literature review research on missing values, the first step is to conduct a comprehensive literature search. This process involves the use of various academic databases, scientific journals, repositories and other reliable sources of information. A literature search should be carried out with relevant keywords covering a specific period of literature to ensure that relevant and up-to-date literature is included in the study.

2.2 Research Problems

After conducting a literature search, studies should apply inclusion and exclusion criteria to select suitable studies for consideration in the review. Inclusion criteria should be clear and specific, such as relevant research fields (eg, social sciences, medicine, economics), type of research (experimental studies, surveys, observational studies, etc.), and primary focus on addressing missing values. On the other hand, exclusion criteria should help filter out studies that are irrelevant or do not meet certain quality standards. In the context of research, PICOC helps researchers to formulate research questions in a structured way by identifying Population, Intervention, Comparison, Outcome to be observed, and the research topic area (Context). This facilitates the search for relevant evidence and ensures clear research questions so that they can be answered in a systematic and objective manner [23]. The research questions underlying this literature review can be seen in table 2.

Table 1. Summary of PICOC.

Population	Data, Value, missing data, missing value,
Intervention	Missing value, missing data, metode, model, dataset
Comparison	n/a
Outcome	The method with the highest accuracy in repairing missing values
Context	Public non-health data set

Table 2 Research Questions on Literature Review

Id	Reseach question	Motivation
RQ1	which journal is the most significant missing value journal?	Identify the most significant journal in the missing value area
RQ 2	Who are the most active researchers in the field of missing values	Identification of researchers who have contributed the most to the missing value research field
RQ 3	What research topics are chosen by researchers in the field of missing values?	Identify topic areas and trend missing values
RQ 4	What datasets are most widely used in missing value research	Identify the dataset used in the missing value
RQ 5	What method is used in missing value research	Identify the opportunities and trends of the methods used in missing values
RQ 6	What method is most widely used in missing value research	Identify the opportunities and trends of the most widely used methods for missing values
RQ 7	Which method performs best when used to correct missing values?	Identify the best method in missing value
RQ 8	What method is proposed to fix the missing value	Identification of the best method proposed to fix the missing value
RQ 9	What framework is proposed to fix missing values	Identify frameworks that can be used to correct missing values

2.3 Search Strategy

Once suitable studies have been selected based on inclusion and exclusion criteria, the next step is to evaluate the literature carefully. Literature evaluation involves reading and critical analysis of each article or publication included in the review. The factors evaluated included the study design, the missing value handling method used, sample size, data validity, and the power of statistical analysis. The digital database of article searches was conducted at sciencedirect.com. The search string is developed according to the following steps:

1. Identify search terms from PICOC, especially from Population and Intervention
2. Identify the search terms from the research question
3. Identify search terms in the title, abstract, and relevant keywords
4. Identify synonyms, alternative spellings, and antonyms of search terms
5. Sophisticated search string construction using identified search terms, Boolean AND and OR.

Here are the search strings used:

(missing) AND (valu OR data*) AND (Imput*) AND (Proces*)* *(1)*

Search string adjustments were made, but the original was kept, because a search string adjustment would dramatically increase an already extensive list of irrelevant studies. The search string is then customized according to the specific needs of each database. The database is searched by title, keyword and abstract. Searches are limited by year of publication: 2010-2023. Two types of publications i.e. journal papers and conference proceedings are included. The search was also limited to articles published in English.

2.4 Selection of Studies

Furthermore, relevant data must be extracted from each selected study. The data extraction process includes important information such as authors, year of publication, research design, methods used to handle missing values, results of analysis, and main findings of each study. This data can then be organized in tables or systems that can assist in subsequent analysis see figure 3. In the context of research or data analysis, the terms inclusion and exclusion refer to the selection process and the selection of criteria used to decide whether an element (for example, participants in a study data, or studies) will be included (inclusion) or excluded (exclusion) in a particular analysis or study. Following are the differences between inclusion and exclusion:

Table 3 Inclusion and Exclusion Criteria

Inclusion Criteria	Studies in academia and industry use both large and small scale data sets
	The study discusses and compares the performance of modeling in the field of software defect prediction
	For studies that have conference and journal versions, only the journal version will be included
	For duplicate publications of the same study, only the most complete and most recent will be included
Exclusion Criteria	Studies without strong validation or include software crash prediction experimental results
	Studies that address defect prediction data sets, methods, frameworks in contexts other than software defect prediction
	Studies are not written in English

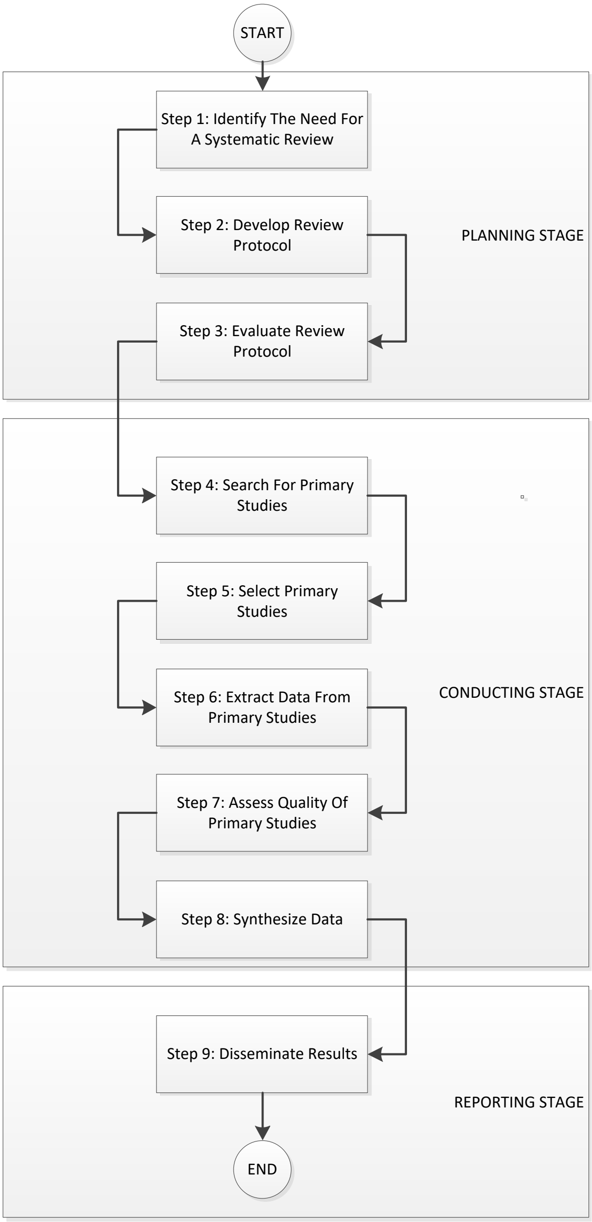

START
Step 1: Identify The Need For A Systematic Review
Step 2: Develop Review Protocol
Step 3: Evaluate Review Protocol
PLANNING STAGE
Step 4: Search For Primary Studies
Step 5: Select Primary Studies
Step 6: Extract Data From Primary Studies
Step 7: Assess Quality Of Primary Studies
Step 8: Synthesize Data
CONDUCTING STAGE
Step 9: Disseminate Results
END
REPORTING STAGE

Fig.1 Systematic Review Stages

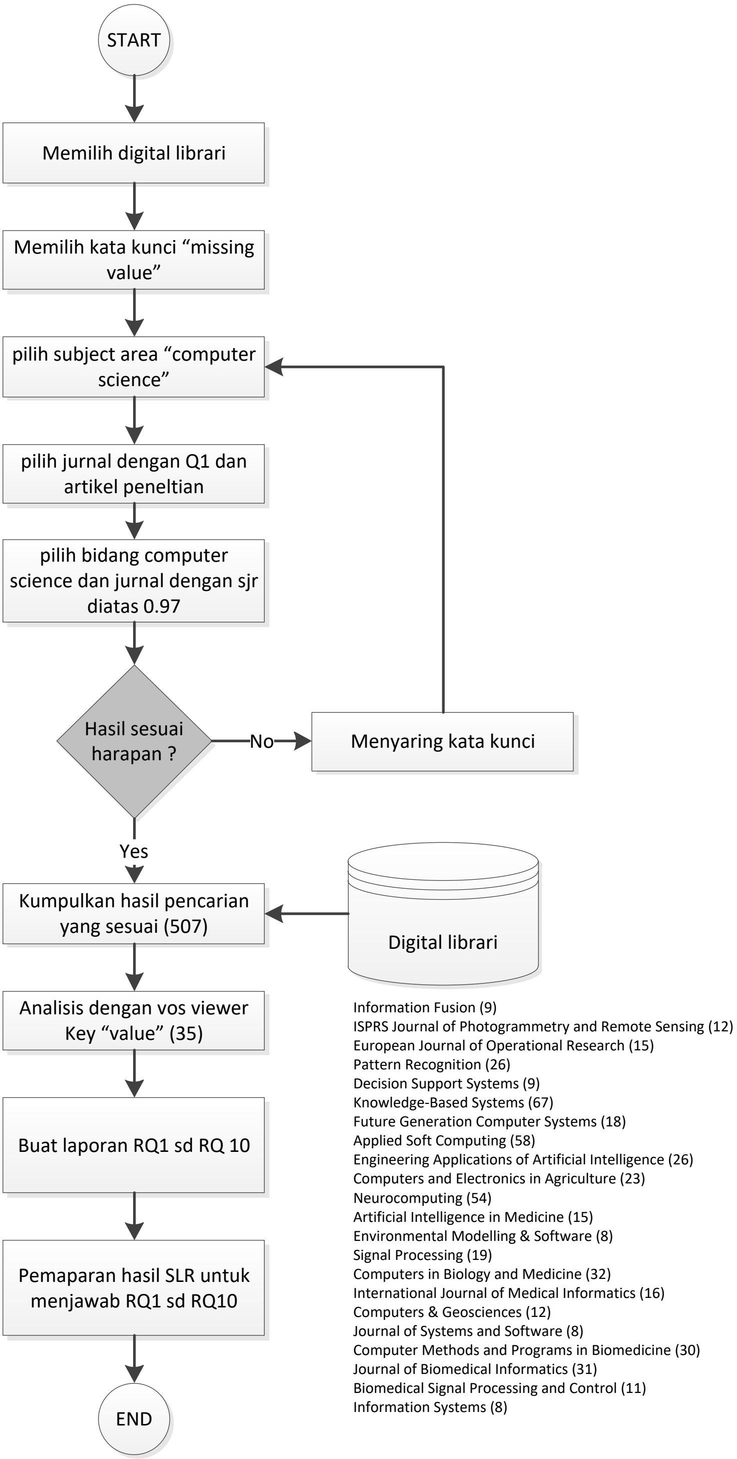

START
Memilih digital librari
Memilih kata kunci “missing value”
pilih subject area “computer science”
pilih jurnal dengan Q1 dan artikel peneltian
pilih bidang computer science dan jurnal dengan sjr diatas 0.97
Hasil sesuai harapan ?
No
Menyaring kata kunci
Yes
Kumpulkan hasil pencarian yang sesuai (507)
Digital librari
Analisis dengan vos viewer Key “value” (35)
Buat laporan RQ1 sd RQ 10
Pemaparan hasil SLR untuk menjawab RQ1 sd RQ10
END
Information Fusion (9)
ISPRS Journal of Photogrammetry and Remote Sensing (12)
European Journal of Operational Research (15)
Pattern Recognition (26)
Decision Support Systems (9)
Knowledge-Based Systems (67)
Future Generation Computer Systems (18)
Applied Soft Computing (58)
Engineering Applications of Artificial Intelligence (26)
Computers and Electronics in Agriculture (23)
Neurocomputing (54)
Artificial Intelligence in Medicine (15)
Environmental Modelling & Software (8)
Signal Processing (19)
Computers in Biology and Medicine (32)
International Journal of Medical Informatics (16)
Computers & Geosciences (12)
Journal of Systems and Software (8)
Computer Methods and Programs in Biomedicine (30)
Journal of Biomedical Informatics (31)
Biomedical Signal Processing and Control (11)
Information Systems (8)

Fig. 2 Search and Selection of Primary Studies

2.5 Data Extraction

During data analysis, the results of the evaluated and extracted studies will be systematically synthesized. This involves compiling findings from each study to identify trends, similarities, differences, and the contributions of each study in the context of handling missing values. This analysis can be done narratively or using certain statistical methods such as meta-analysis, if there is sufficient data to be carried out.

Table 2. Summary of PICOC.

Property	Research questions
Researchers and Publications	RQ1, RQ2
Research Trends and Topics	RQ3
Missing value Datasets	RQ4
matrik tool yang dipakai	RQ4
Misiing value Prediction Methods	RQ5, RQ6, RQ7, RQ8
missing value Frameworks	RQ9

2.6 Assessment of Study Quality and Data Synthesis

After the data analysis is complete, the researcher must interpret the findings of the review. This interpretation must be based on the evidence available in the literature and must draw strong and credible conclusions about effective methods of dealing with missing values.

3 Research Result

3.1 Which Journal Is The Most Significant Missing Value Prediction Journal

In this literature review, 33 selected articles were extracted and answered the research questions included in the protocol. The distribution of articles by year can be seen in that missing value research has been popular since 2013. Then it decreased in the following 2 years and experienced an increase again in the following 2 years. This article was obtained from science direct. It is possible that other reputable journals do not experience this. Reputable journals are scientific journals that have high recognition and credibility in the academic world. Usually, the assessment of journal reputation is based on factors such as impact factors, number of citations, quality of published articles, and recognition from the scientific community. Some of the indexes of reputable journals that are commonly used are: Web of Science (WoS), Scopus, PubMed, IEEE Xplore, Google Scholar, Directory of Open Access Journals (DOAJ), Emerging Sources Citation Index (ESCI), ERIC (Education Resources Information Center) , PsycINFO, and the Chemical Abstracts Service (CAS).

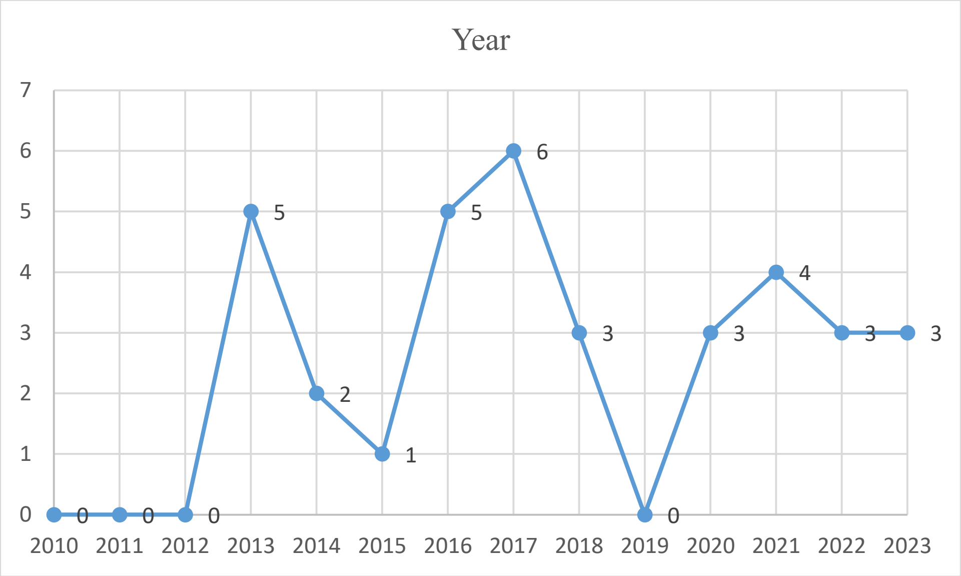


Fig. 3 Distribution of articles by year

In figure 4 the journal that publishes the most articles about missing value is Knowledge-Based Systems, this journal has the scope of knowledge-based and other artificial intelligence techniques-based systems with the following objectives and capabilities: to support human prediction and decision-making through data science and computation techniques; to provide a balanced coverage of both theory and practical study in the field; and to encourage new development and implementation of knowledge-based intelligence models, methods, systems, and software tools, with applications in business, government, education, engineering and healthcare. This journal also has a fairly short publication time, which is around 1.3 weeks. As shown in table 1, the value of the Scientific Journal Rankings (SJR) for the journal Knowledge-Based Systems 2.07 is quite high. The sources of information we get are the last 10 years (2013-2023). In the Knowledge-Based Systems journal, there are 67 articles with missing value topics based on title, abstract and keyword searches conducted on sciendirect. The highest research occurred in 2018, 2019 and 2022. It had decreased in 2021 with 7 articles.

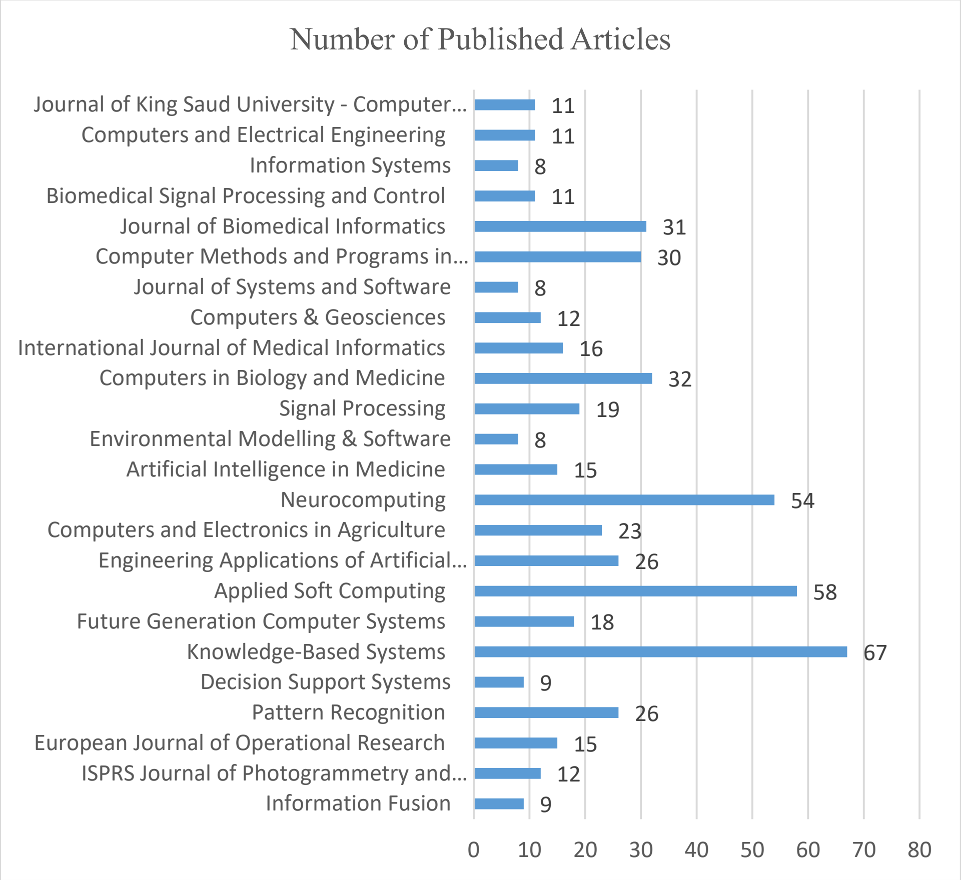


Fig. 4 Journal Publications and Distribution of Selected Studies

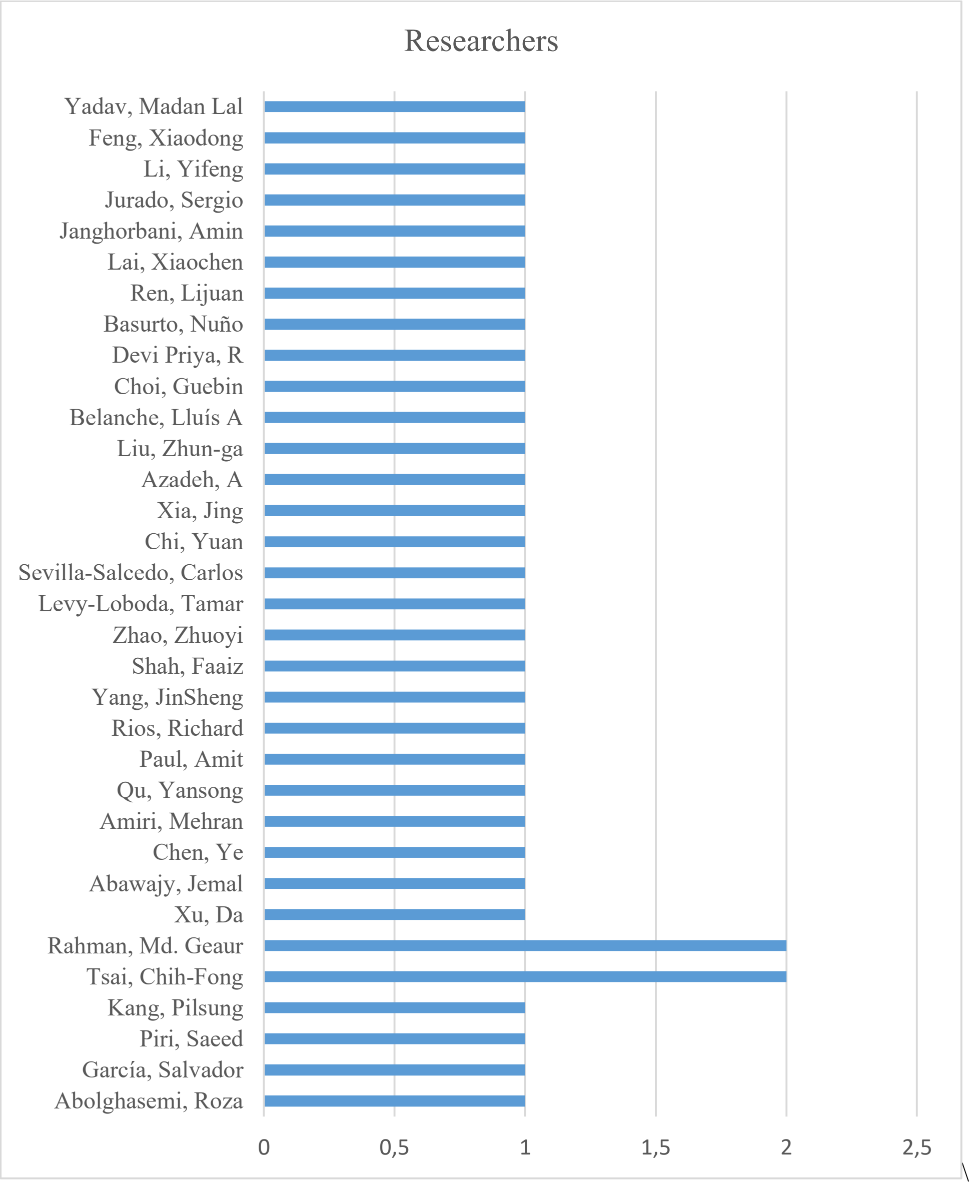
Table 5. Journal Publications and Distribution of Selected Studies

Journal	Quantity	SJR	Country
Information Fusion	9	4.76	Netherland
ISPRS Journal of Photogrammetry and Remote Sensing	12	3.31	Netherland
European Journal of Operational Research	15	2.37	Netherland
Pattern Recognition	26	2.09	United Kingdom
Decision Support Systems	9	2.08	Netherland
Knowledge-Based Systems	67	2.07	Netherland
Future Generation Computer Systems	18	2.04	Netherland

Applied Soft Computing	58	1.88	Netherland
Engineering Applications of Artificial Intelligence	26	1.73	United Kingdom
Computers and Electronics in Agriculture	23	1.59	Netherland
Neurocomputing	54	1.48	Netherland
Artificial Intelligence in Medicine	15	1.44	Netherland
Environmental Modelling & Software	8	1.35	Netherland
Signal Processing	19	1.23	Netherland
Computers in Biology and Medicine	32	1.22	United Kingdom
International Journal of Medical Informatics	16	1.2	Irlandia
Computers & Geosciences	12	1.18	United Kingdom
Journal of Systems and Software	8	1.13	United States
Computer Methods and Programs in Biomedicine	30	1.12	Irlandia
Journal of Biomedical Informatics	31	1.08	United States
Biomedical Signal Processing and Control	11	1.07	Netherland
Information Systems	8	0.98	United Kingdom

3.2 Who Are The Most Active And Influential Researchers In The Field Of Missing Value Repair

Based on the last 10 years of research on missing values, data can be drawn regarding which authors or researchers are the most active. Uniquely, one of the most active researchers on the topic of missing values comes from Indonesia with the title Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques [15] and [24]. The second most are researchers from China with the title Combining instance selection for better missing value imputation [25] and A class center based approach for missing value imputation [3]. When pulled back 5 years the missing value has been studied by Lijuan Ren, JinSheng Yang, Yansong Qu, Richard Rios, Roza Abolghasemi, Carlos Sevilla-Salcedo, Xiaochen Lai, Zhuoyi Zhao, Nuño Basurto, Tamar Levy-Loboda, Saeed Piri, Da Xu, Faaiz Shah, ChihFong Tsai, MadanLal Yadav, Guebin Choi.



\

Fig. 5 Influential Researchers and Number of Studies

3.3 What Kind Of Research Topics Are Chosen By Researchers In The Field Of Missing Value Repair

Withdrawing information related to the research topic was carried out by selecting 10 papers that had been included in the protocol. Papers [15], [25]–[33] are included in the 10 studies with the highest number of citations. Missing value research is categorized into 5 topics [27] :

1. Estimation
2. Association
3. Classification

4. Clustering
5. Dataset Analysis

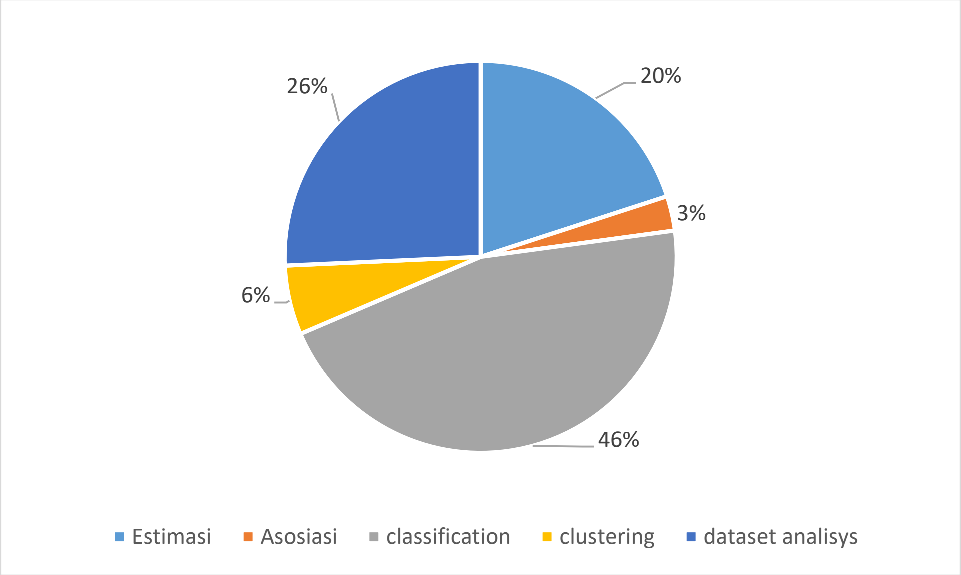


Fig. 6 Distribution of Research Topics

The first type, namely estimation, has been carried out by [34] showing empirical results on various sensor data showing the superiority of the proposed framework, especially in overcoming large segment imputation problems, as evidenced by increased imputation performance. Lijuan Ren [2] refined the data collection process as early withdrawal and participant rejection, as a way to resolve where there was a lot of missing value in medical data. Based on this embedding, another novelty of the approach taken regarding missing values is using the similarities of experts, as well as similarities between alternatives, to infer missing values even when only minimal data is available for several alternatives from several experts [10]. The results show that the eight strategies proposed are able to overcome MV and utilize the information embedded in the data, with better results in those strategies that utilize causal relevance [35]. Experimental results show that GL2P outperforms its competitors in terms of imputation accuracy and better maintains the structure of differentially expressed genes [36]. The proposed LCSR-MVE outperforms other state-of-the-art methods in terms of normalized root mean square error (NRMSE), and is not very sensitive to dictionary size and regularization parameters [37]. The computational results show that the best answer (in the population of the last 10 chromosomes) obtained by GA is often the same as the missing value, with the average value approaching the missing observations [38].

The second type is association; this study has the least amount compared to the others. The proposed Dual Repopulated Bayesian Ant Colony Optimization (DPBACO) handles non-negligible and non-negligible missing values in heterogeneous attributes of large data sets. DPBACO integrates Bayesian principles with Ant Colony Optimization techniques because they are both simple and efficient to implement. After pheromone updating, repopulation of the solution pool is carried out by dividing the population into two based on its fitness value and generating new offspring by carrying out a crossover operation [39]. Not many articles are included in the SLR protocol that we made. However, if it is carried out outside the protocol, there are still several similar studies such as [40] where methods used on non-clinical data containing chemical-disease associations to find relationships between different phenotypes, such as prostate cancer and breast cancer can be improved and show good results. good at repairing missing values.

The third type is classification where this type has the most number of researchers. the proposed SSHIBA framework can learn excellent imputation of missing values and outperform the baseline while simultaneously predicting three different tasks [41]. In the conclusion of this third type of research, it can be concluded that the classification approach has proven effective in overcoming the problem of missing values in various contexts [41][42][43][44]. Through the application of classification algorithms such as K-Nearest Neighbors, Naive Bayes, Random Forest, SVM, and ensemble models, we have succeeded in developing solutions that can accurately predict or classify data even if there are missing values [45][46][47][7][8][48]. In many cases, this approach can improve the performance of classification models as well as provide valuable insights into data analysis. However, it should be remembered that the selection of the right method, a good understanding of the data, and careful evaluation of the imputed results are very important to ensure that the results obtained are reliable and relevant in a practical context [49][25][27][5][13][50].

The fourth type of research is clustering. This study only has 2 articles that are included in the slr protocol that we made. This study evaluates four real-world data sets and twenty basic models. Experimental results show that the proposed ST-A-PGCL achieves superior predictive performance, especially in long-term prediction tasks with high loss rates i [51]. Whereas what was done by [52] showed that missing values reduce the accuracy of the ML model when predicting MACE risk. Removing variables with missing values and retraining the model can lead to superior patient-grade predictive performance.

The last type, namely the analysis dataset, ranks second with the number of articles 9. In this study, it shows success in overcoming challenges that arise due to missing values in the analysis dataset [53]. By applying an approach that focuses on classification techniques, we are able to fill in missing values with a sufficient degree of accuracy. This approach proves itself as a reliable solution in dealing with incomplete data problems [3]. The results show that classification methods, such as K-Nearest Neighbors, Naive Bayes, and other algorithms, can effectively predict or classify data that contains missing values [54]. This allows us to take full advantage of the information available in the dataset, so that the analysis results are more accurate and meaningful [55]. However, the success of this approach cannot be ignored by selecting the right model, deep understanding of the characteristics of the dataset, and careful evaluation of the resulting model performance [56]. Although the classification approach has shown promising results, the sustainability of this method in the context of data analysis will depend on its ability to cope with more complex variations in data that approximate real-world conditions [57]. Thus, this study provides valuable insights into how a classification approach can be used effectively to fill in missing values in analytical datasets [58][24]. However, broader challenges and more complex cases still need further research to develop a stronger and more reliable approach in handling incomplete data [15].

3.4 What Type Of Dataset Is Most Widely Used For Repairing Missing Values

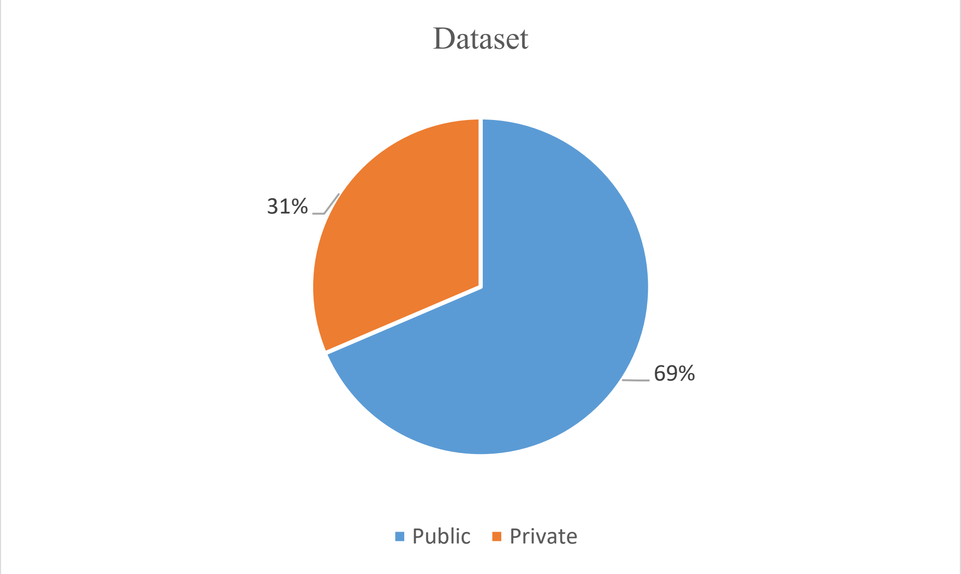


Fig. 7 Total Distribution of Datasets

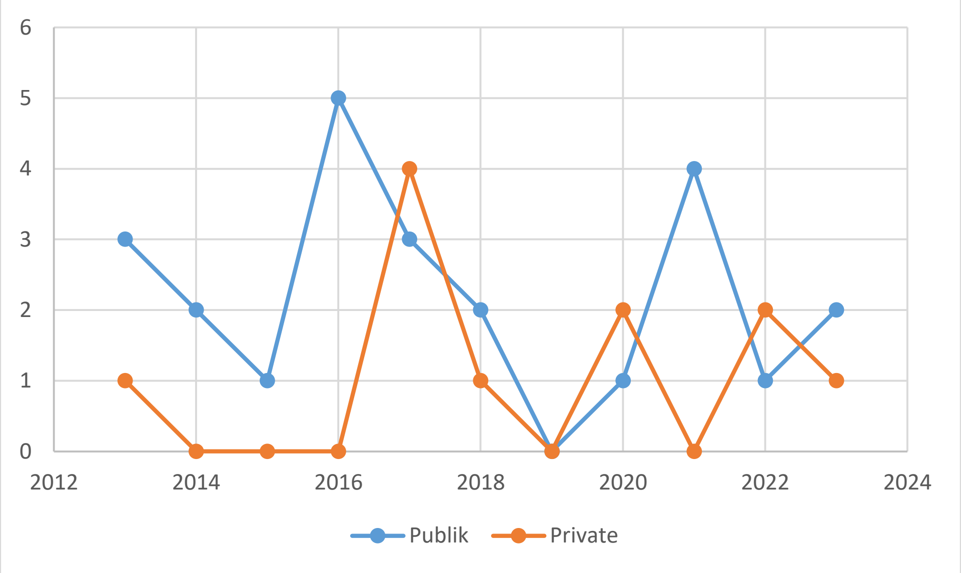


Fig. 8 Distribution of Private and Public Datasets

Datasets play a key role in research related to missing values. Dataset is a collection of data used as a basis for analysis and research. In the context of missing value research, datasets have an important role in several aspects:

1. Research Context: The dataset determines the context and purpose of the research. An in-depth understanding of the dataset helps researchers identify patterns and trends that may be affected by missing values, and decide on the most appropriate approach to addressing the problem.
2. Identification of Missing Values: Datasets are the main source of information about where and how much of the missing values are in the data. Researchers can analyze

the dataset to identify patterns of missing values, whether the missing values appear randomly or have a certain structure.

3. Selection of Approach: The dataset influences the choice of approach to be used to address missing values. The characteristics of the dataset, such as the number and type of variables, the number of missing values, and the relationship between variables, will influence the choice of the right imputation method.
4. Evaluation of Results: The dataset is used to evaluate the effectiveness of the approach applied in filling in the missing value. A comparison between the original data and the imputed data provides insight into how well the approach is performing.
5. Further Analysis: Datasets that have been imputed or that have resolved missing values can be used for further analysis. This allows researchers to gain new insights or make predictions based on exhaustive data.
6. Research Reproduction: The dataset obtained and the resulting imputation results must be maintained so that the research can be reproduced by others. This ensures the validity and transparency of the research in addressing the problem of missing values.
7. Additional Research: The processed dataset can be used to conduct additional research or further modelling. More complete data allows researchers to explore new questions or develop more accurate models.

As shown in Figure 7 regarding the number of datasets used on the topic of missing values over the past 10 years, 69% of public datasets were used, followed by 31% private datasets. Figure 8 shows the distribution of dataset usage in missing value research from year to year for the last 10 years (2013-2023). Most public datasets come from machine learning uci. Although there are some private datasets that are later made public for scientific purposes. There are several private data sets that will be provided if the researcher is interested in researching the same topic by contacting the contact of the main researcher.

3.5 What Type Of Method Is Used To Correct Missing Values

In data analysis research, the presence of missing values can be a significant obstacle in producing accurate and meaningful results. Therefore, the adoption of effective methods to address these issues is essential in ensuring the integrity of the analysis. Various approaches have been developed to fill in or overcome missing values in datasets. These methods involve techniques from various disciplines, such as statistics, machine learning, and relevant knowledge domains. Selection of the right method depends on the characteristics of the dataset, the distribution of missing values, and the ultimate goal of the analysis. In the following discussion, we will explore some common methods used in missing value research, including imputation using the average value, median, or mode, approaches based on machine learning algorithms such as K-Nearest Neighbors (KNN) [2], [35], [44], [45], [51] and Random Forests [2], [3], [56], as well as sophisticated statistical methods such as Multiple Imputation. By understanding and applying the appropriate methods, we can generate more complete datasets and allow for more in-depth analysis and richer information. Figure 9 shows the methods used by researchers in the last 10 years.

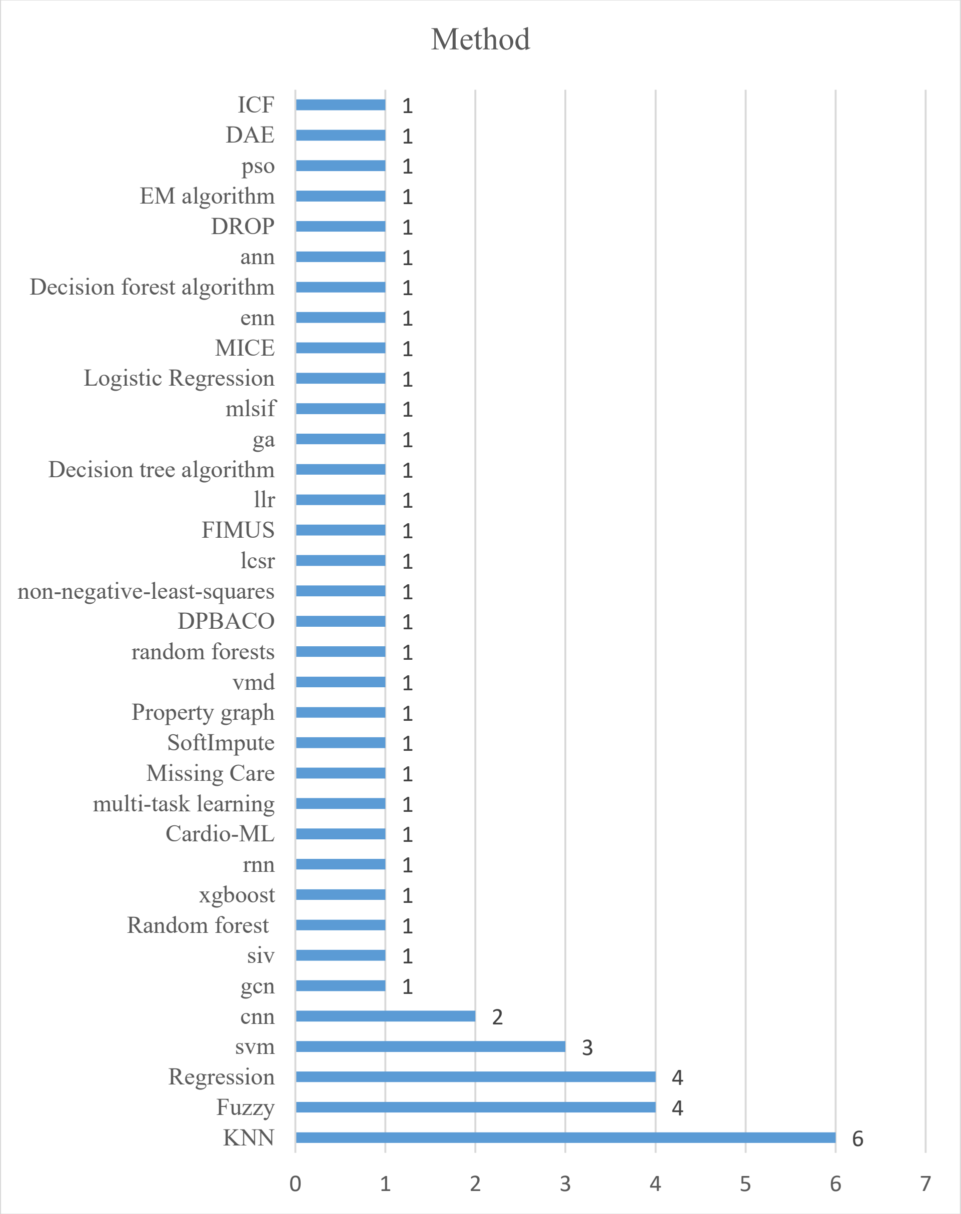


Fig. 9 Methods used in Missing Value

3.6 What Type Of Method Is Most Often Used To Correct Missing Values

In an effort to overcome the missing value problem, the imputation method is the most common option that is often used. This method involves filling in missing values based on the information contained in the dataset. In the category of imputation methods, the most frequently used is imputation using the average or median value of the related variables. In addition, imputation using the K-Nearest Neighbors (KNN) algorithm is also often relied upon, because this method utilizes information from nearest neighbors in filling in missing values. Regression methods, both linear and non-linear, are also often applied to fill in missing values by utilizing the relationship between the relevant variables. In each choice of imputation method, it is necessary to carry out evaluation and validation to ensure that the results obtained meet the desired analytical objectives.

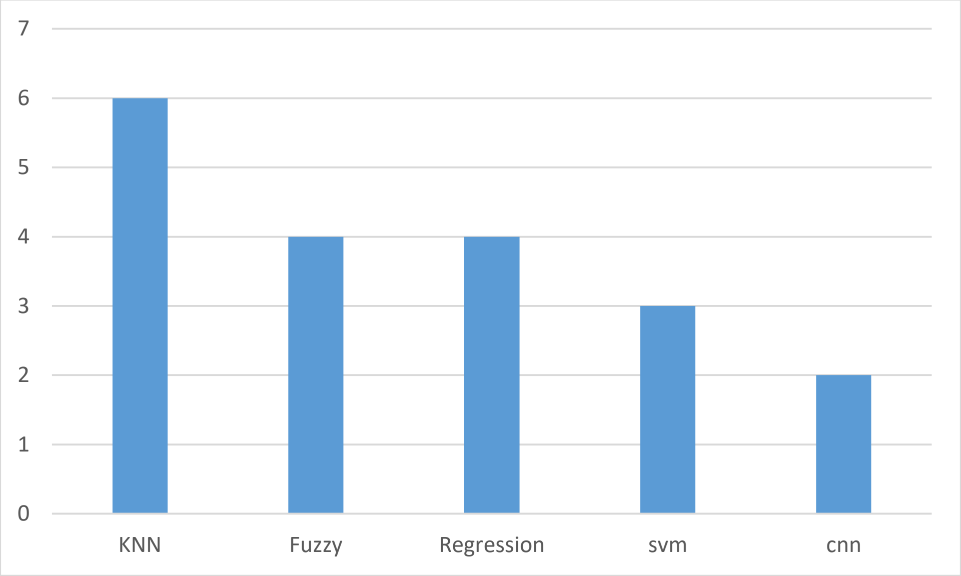


Fig. 10 Most Used Methods in missing value

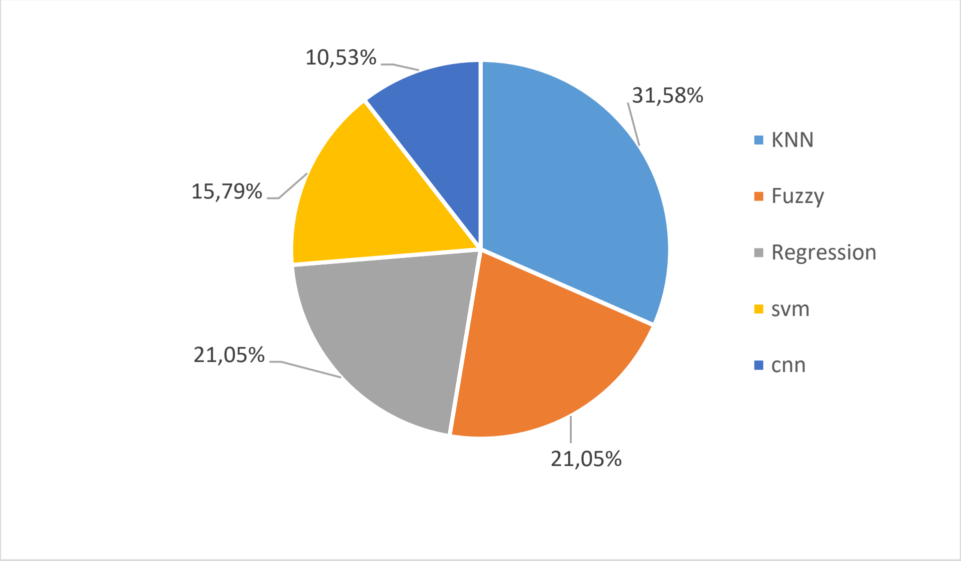


Fig. 11 Distribution of the Studies over Type of Methods

Another imputation method that is often used is filling by using the mode (the value that appears the most) for categorical variables. This method is effective in overcoming missing values in variables with discrete or categorical data types. In addition, imputation with machine learning algorithms such as Random Forests or methods based on deep learning is also gaining popularity. The advantage of this method lies in its ability to capture complex relationships between variables, which can provide better results in filling in missing values. Figure 10 shows the use of KNN still dominates over other methods with a total of 6 articles.

In practice, the choice of imputation method will be greatly influenced by the characteristics of the data and the type of analysis to be performed. Some methods are better suited for data with a certain structure, while others are better suited for more complex datasets. Therefore, the selection of the most appropriate imputation method must be based on a deep understanding of the dataset and the research objectives to be achieved. Thus, choosing the appropriate imputation method can make a significant contribution in ensuring the integrity and accuracy of the analysis results. Figure 11 shows the percentage of the most widely used method in missing value research in the last 10 years from the digital sciendirect database.

3.7 Which Method Performs Best When Used For Missing Value Repair

Many research activities have carried out missing value repairs with several algorithms. Some of the algorithms used can be seen in the table. There is no strong consensus regarding which algorithm or method is best for repairing missing values when research is viewed individually [57]. No single method is absolutely the best performer in all cases when used for missing value repair. The effectiveness of these methods is highly dependent on the nature and characteristics of the dataset, including the number and distribution of missing values, the types of variables, and the relationships between these variables. In some situations, imputation using the mean or median values can give satisfactory results, especially if the missing values are randomly distributed. However, if there is a more complex structure in the data, methods such as K-Nearest Neighbors, machine learning algorithms or regression may be more effective. Evaluation of the performance of these methods should be carried out through cross-validation or other techniques to ensure that the imputations produced are suitable for the purpose of analysis and provide accurate and consistent results.

In addition, the use of multiple imputation methods can also provide more accurate results in conditions of high uncertainty or when the dataset has a complex structure. In some cases, a combined approach, where several methods are used together to fill in missing values, may produce better results than a single method. Even so, it cannot be avoided that each method has its own limitations and assumptions that need to be considered. In making decisions about the method to be used, it is important to consider the purpose of the analysis, the characteristics of the dataset, and the available computing capabilities. It is also good practice to experiment with different methods and compare the imputation results and their impact on further analysis to assess the performance of the methods. With a deep understanding of the various methods and their compatibility with certain situations, researchers can choose and apply the most suitable missing value repair approach to achieve optimal results.

3.8 What Type Of Repair Method Is Proposed For Missing Value Repair

In an effort to overcome the challenge of missing values, various innovative approaches have been proposed that combine various methods. One such approach involves the use of the Feature Selection technique as shown in Figure 12, in which the most relevant features are selected for imputation, thereby minimizing the impact of missing values on the analysis results [49]. A mixed approach (mix method) has also been applied, combining statistically based imputation with machine learning algorithms to take advantage of the strengths of both [2], [59].

In addition, Ensemble Machine Learning, which combines several models to produce stronger predictions, has become the main focus in overcoming missing values [3], [4], [45], [49], [56], [57]. This approach combines different machine learning models to fill in missing values, increasing accuracy and robustness to data variation. Furthermore, a framework that unifies various methods in a structured framework also emerges as the latest approach in repairing missing values. This framework provides guidance for researchers in selecting the most suitable method based on the nature of the dataset and the purpose of the analysis[3], [9], [56], [60].

The use of Feature Selection in the context of missing value correction is very important because it helps reduce data dimensions by focusing on the most informative features, resulting in more accurate and efficient imputation. The mixed method combines the advantages of various methods, such as statistically based imputation and machine learning algorithms, to produce a more holistic solution to the problem of incomplete data. Ensemble Machine Learning shows great potential in overcoming the problem of missing values because it leverages the power of various models to produce more robust and stable imputation. This approach does not rely solely on one model, but incorporates multiple perspectives, so that it is able to handle variations and complexities in datasets. Meanwhile, the application of the framework provides a structured guide for researchers in selecting and applying appropriate methods in overcoming missing values, thereby minimizing the risk of errors in method selection.

Overall, the approach that combines Feature Selection, blended approaches, Ensemble Machine Learning, and utilization of frameworks marks an advance in handling missing values. Through this approach, research does not only focus on filling in missing values, but also on developing comprehensive and effective solutions to increase the integrity and value of information obtained from data analysis

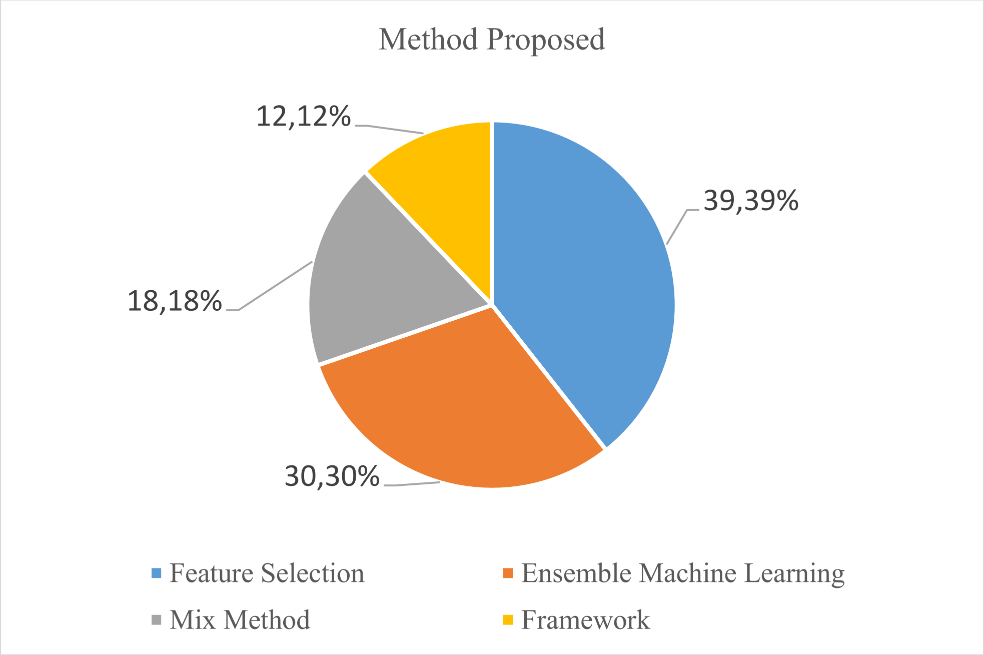


Fig. 12 The most widely used method in repairing missing values

3.9 What Type Of Framework Is Proposed For Missing Value Repair

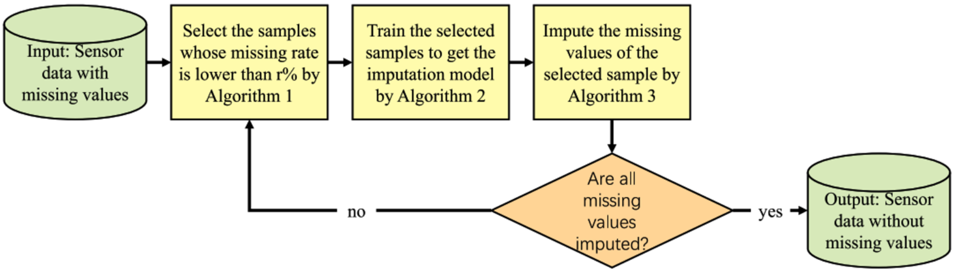

Input: Sensor data with missing values
Select the samples whose missing rate is lower than r% by Algorithm 1
Train the selected samples to get the imputation model by Algorithm 2
Impute the missing values of the selected sample by Algorithm 3
Are all missing values imputed?
no
yes
Output: Sensor data without missing values

Fig.13 Multistage Deep Imputation Framework

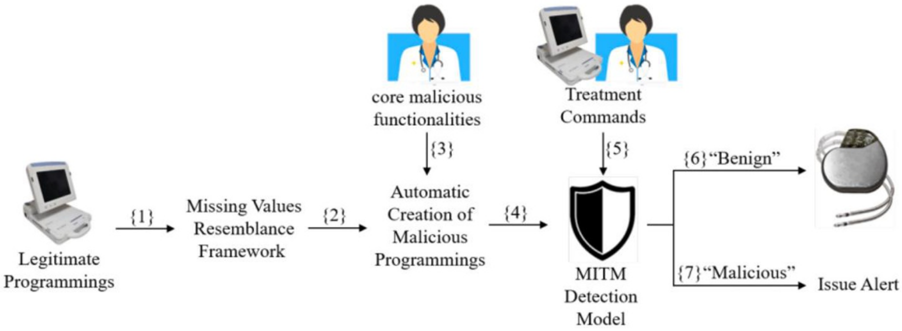


Fig. 14 Cardio-ML System Framework

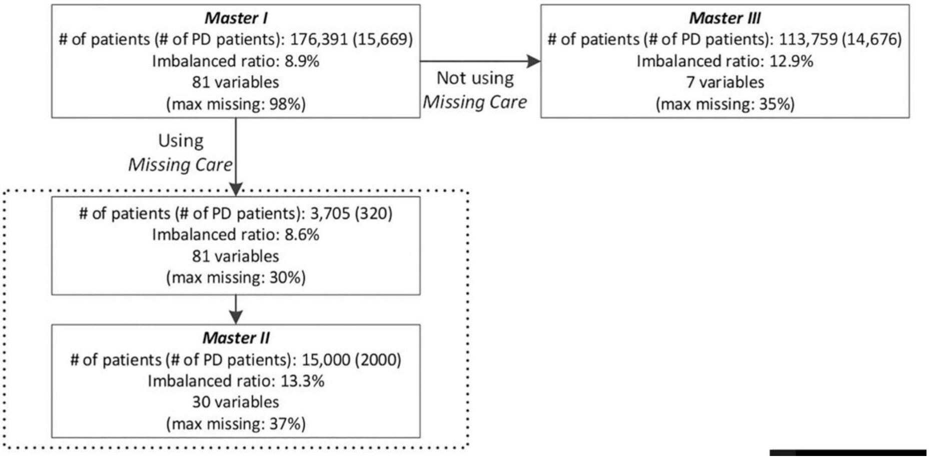


Fig. 15 Missing Care Framework

The Multistage Deep Imputation Framework (see figure 13) is an approach that involves several stages in overcoming missing values [53]. In the first stage, relevant features are selected using the Feature Selection technique to minimize the impact of missing values. Furthermore, statistical imputation or regression methods are used to fill in some of the missing values. In the next stage, Deep Learning models, such as Autoencoders, are used for further imputation by utilizing the information contained in the imputed data and features. This results in more accurate and comprehensive imputation results. This approach combines the strengths of multiple methods in one structured framework, ultimately resulting in more robust imputation and increased quality of data analysis.

The Cardio-ML System Framework is a framework that consists of several steps in the analysis of cardiovascular data. First, clinical data is collected and pre-processed to address missing values and outliers. Next, important features were extracted using a feature selection technique based on clinical significance. After that, various machine learning algorithms such as Random Forest, Support Vector Machines, and Neural Networks were applied to build predictive models capable of classifying cardiovascular risk. Model results are evaluated through cross-validation and performance evaluation metrics. Furthermore, the predictive results and insights from the model can assist in making more informed and efficient clinical decisions in the management of cardiovascular disease (see figure 14).

The Missing Care Framework is an approach that identifies and analyzes medical actions that are not performed or are missed in patient care. First, patient care data is collected and recorded in detail. Furthermore, the data is analyzed to identify medical actions that should be performed but are not recorded in the medical records. Then, the measures were categorized by type and level of clinical importance. From this, insights are obtained about the tendency of medical action to be missed and the factors that influence it. This information can be used to improve the quality of patient care and identify areas where improvements in the provision of care are needed (see figure 15).

4 Conclusions and Future Work

In conclusion, a literature review related to the topic of missing values in the last 10 years obtained from Sciencedirect shows a significant evolution in the development of methods and approaches to addressing the problem of missing values. Various imputation methods, such as statistically based imputation, machine learning algorithms, and mixed approaches, have been applied to improve the integrity and quality of data analysis. Innovative approaches, such as the use of machine learning-based Deep Learning and Frameworks, are increasingly demonstrating the potential to more accurately and effectively address the challenge of incomplete data.

However, although many studies have contributed greatly to this field, there are still some aspects that need attention. First, further research can explore the potential of ensemble methods, such as combining various imputation algorithms, to produce more consistent and robust results. Second, more stringent validation and evaluation of the imputed results is needed to ensure the reliability of the resulting solutions. Third, research can also further explore the application of the imputation method in various domains, such as health, finance, or social sciences. Finally, a focus on developing frameworks or systems that can guide practitioners in selecting appropriate methods will be invaluable in addressing the complexity and variation in real use cases.

Overall, literature reviews in the last 10 years from Sciencedirect have provided a comprehensive view of the latest developments in addressing the problem of missing values. Future research is expected to continue to enrich and develop better solutions to address this still relevant problem.

References

[1] C. Sevilla-Salcedo, V. Imani, P. M. Olmos, V. Gómez-Verdejo, and J. Tohka, “Multi-task longitudinal forecasting with missing values on Alzheimer’s disease,” *Comput. Methods Programs Biomed.*, vol. 226, p. 107056, 2022, doi: 10.1016/j.cmpb.2022.107056.

[2] L. Ren, A. S. Seklouli, H. Zhang, T. Wang, and A. Bouras, “An adaptive Laplacian weight random forest imputation for imbalance and mixed-type data,” *Inf. Syst.*,

vol. 111, p. 102122, 2023, doi: https://doi.org/10.1016/j.is.2022.102122.

[3] C. F. Tsai, M. L. Li, and W. C. Lin, “A class center based approach for missing value imputation,” *Knowledge-Based Syst.*, vol. 151, pp. 124–135, 2018, doi: 10.1016/j.knosys.2018.03.026.

[4] Z. G. Liu, Q. Pan, J. Dezert, and A. Martin, “Adaptive imputation of missing values for incomplete pattern classification,” *Pattern Recognit.*, vol. 52, pp. 85–95, 2016, doi: 10.1016/j.patcog.2015.10.001.

[5] L. A. Belanche, V. Kobayashi, and T. Aluja, “Handling missing values in kernel methods with application to microbiology data,” *Neurocomputing*, vol. 141, pp. 110–116, 2014, doi: https://doi.org/10.1016/j.neucom.2014.01.047.

[6] M. G. Rahman and M. Z. Islam, “Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques,” *Knowledge-Based Syst.*, vol. 53, pp. 51–65, 2013, doi: 10.1016/j.knosys.2013.08.023.

[7] A. Paul, J. Sil, and C. Das Mukhopadhyay, “Gene selection for designing optimal fuzzy rule base classifier by estimating missing value,” *Appl. Soft Comput.*, vol. 55, pp. 276–288, 2017, doi: https://doi.org/10.1016/j.asoc.2017.01.046.

[8] Y. Chi, J. Hong, A. Jurek, W. Liu, and D. O’Reilly, “Privacy preserving record linkage in the presence of missing values,” *Inf. Syst.*, vol. 71, pp. 199–210, 2017, doi: https://doi.org/10.1016/j.is.2017.07.001.

[9] A. Azadeh *et al.*, “Optimum estimation of missing values in randomized complete block design by genetic algorithm,” *Knowledge-Based Syst.*, vol. 37, pp. 37–47, 2013, doi: 10.1016/j.knosys.2012.06.014.

[10] R. Abolghasemi, R. Khadka, P. G. Lind, P. Engelstad, E. H. Viedma, and A. Yazidi, “Predicting missing pairwise preferences from similarity features in group decision making,” *Knowledge-Based Syst.*, vol. 256, p. 109860, 2022, doi: https://doi.org/10.1016/j.knosys.2022.109860.

[11] M.-C. Wang, C.-F. Tsai, and W.-C. Lin, “Towards missing electric power data imputation for energy management systems,” *Expert Syst. Appl.*, vol. 174, p. 114743, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114743.

[12] H. Wen, P. Pinson, J. Gu, and Z. Jin, “Wind energy forecasting with missing values within a fully conditional specification framework,” *Int. J. Forecast.*, 2023, doi: https://doi.org/10.1016/j.ijforecast.2022.12.006.

[13] P. Kang, “Locally linear reconstruction based missing value imputation for supervised learning,” *Neurocomputing*, vol. 118, pp. 65–78, 2013, doi: 10.1016/j.neucom.2013.02.016.

[14] H. V. Bhagat and M. Singh, “DPCF: A framework for imputing missing values and clustering data in drug discovery process,” *Chemom. Intell. Lab. Syst.*, vol. 231, p. 104686, 2022, doi: https://doi.org/10.1016/j.chemolab.2022.104686.

[15] M. G. Rahman and M. Z. Islam, “Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques,” *Knowledge-Based Syst.*, vol. 53, pp. 51–65, 2013, doi: https://doi.org/10.1016/j.knosys.2013.08.023.

[16] Y. Shibuya, A. Hamm, and T. Cerratto Pargman, “Mapping HCI research methods for studying social media interaction: A systematic literature review,” *Comput. Human Behav.*, vol. 129, p. 107131, 2022, doi: https://doi.org/10.1016/j.chb.2021.107131.

[17] S. Farivar, F. Wang, and O. Turel, “Followers’ problematic engagement with influencers on social media: An attachment theory perspective,” *Comput. Human Behav.*, vol. 133, p. 107288, 2022, doi: https://doi.org/10.1016/j.chb.2022.107288.

[18] J. R. Ziolkowska, “Economic value of environmental and weather information for

agricultural decisions – A case study for Oklahoma Mesonet," *Agric. Ecosyst. Environ.*, vol. 265, pp. 503–512, 2018, doi: https://doi.org/10.1016/j.agee.2018.07.008.

[19] W. K. Dumenu, "What are we missing? Economic value of an urban forest in Ghana," *Ecosyst. Serv.*, vol. 5, pp. 137–142, 2013, doi: https://doi.org/10.1016/j.ecoser.2013.07.001.

[20] J. Hersch, F. Mendoza Lopez, and J. B. Shinall, "Estimating years of education using the current population survey after 2014," *Econ. Lett.*, vol. 189, p. 109058, 2020, doi: https://doi.org/10.1016/j.econlet.2020.109058.

[21] J. Du and L. Zhou, "Improving financial data quality using ontologies," *Decis. Support Syst.*, vol. 54, no. 1, pp. 76–86, 2012, doi: https://doi.org/10.1016/j.dss.2012.04.016.

[22] C.-H. Cheng, C.-P. Chan, and Y.-J. Sheu, "A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction," *Eng. Appl. Artif. Intell.*, vol. 81, pp. 283–299, 2019, doi: https://doi.org/10.1016/j.engappai.2019.03.003.

[23] L. Ren, T. Wang, A. Sekhari Seklouli, H. Zhang, and A. Bouras, "A review on missing values for main challenges and methods," *Inf. Syst.*, p. 102268, 2023, doi: https://doi.org/10.1016/j.is.2023.102268.

[24] M. G. Rahman and M. Z. Islam, "FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis," *Knowledge-Based Syst.*, vol. 56, pp. 311–327, 2014, doi: 10.1016/j.knosys.2013.12.005.

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

[26] C.-F. Tsai, M.-L. Li, and W.-C. Lin, "A class center based approach for missing value imputation," *Knowledge-Based Syst.*, vol. 151, pp. 124–135, 2018, doi: https://doi.org/10.1016/j.knosys.2018.03.026.

[27] Z. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," *Pattern Recognit.*, vol. 52, pp. 85–95, 2016, doi: https://doi.org/10.1016/j.patcog.2015.10.001.

[28] J. Xia *et al.*, "Adjusted weight voting algorithm for random forests in handling missing values," *Pattern Recognit.*, vol. 69, pp. 52–60, 2017, doi: https://doi.org/10.1016/j.patcog.2017.04.005.

[29] M. G. Rahman and M. Z. Islam, "FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis," *Knowledge-Based Syst.*, vol. 56, pp. 311–327, 2014, doi: https://doi.org/10.1016/j.knosys.2013.12.005.

[30] M. L. Yadav and B. Roychoudhury, "Handling missing values: A study of popular imputation packages in R," *Knowledge-Based Syst.*, vol. 160, pp. 104–118, 2018, doi: https://doi.org/10.1016/j.knosys.2018.06.012.

[31] P. Kang, "Locally linear reconstruction based missing value imputation for supervised learning," *Neurocomputing*, vol. 118, pp. 65–78, 2013, doi: https://doi.org/10.1016/j.neucom.2013.02.016.

[32] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing*, vol. 205, pp. 152–164, 2016, doi: https://doi.org/10.1016/j.neucom.2016.04.015.

[33] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016, doi: https://doi.org/10.1016/j.knosys.2015.12.006.

[34] J. Yang, Y. Shao, C. Li, and W. Wang, "A multistage deep imputation framework

for missing values large segment imputation with statistical metrics," *Appl. Soft Comput.*, p. 110654, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110654.

[35] S. Jurado, À. Nebot, F. Mugica, and M. Mihaylov, "Fuzzy inductive reasoning forecasting strategies able to cope with missing data: A smart grid application," *Appl. Soft Comput.*, vol. 51, pp. 225–238, 2017, doi: https://doi.org/10.1016/j.asoc.2016.11.040.

[36] Y. Chen *et al.*, "A global learning with local preservation method for microarray data imputation," *Comput. Biol. Med.*, vol. 77, pp. 76–89, 2016, doi: https://doi.org/10.1016/j.compbiomed.2016.08.005.

[37] X. Feng, S. Wu, J. Srivastava, and P. Desikan, "Automatic instance selection via locality constrained sparse representation for missing value estimation," *Knowledge-Based Syst.*, vol. 85, pp. 210–223, 2015, doi: https://doi.org/10.1016/j.knosys.2015.05.007.

[38] A. Azadeh *et al.*, "Optimum estimation of missing values in randomized complete block design by genetic algorithm," *Knowledge-Based Syst.*, vol. 37, pp. 37–47, 2013, doi: https://doi.org/10.1016/j.knosys.2012.06.014.

[39] R. Devi Priya, R. Sivaraj, and N. Sasi Priyaa, "Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases," *Knowledge-Based Syst.*, vol. 133, pp. 107–121, 2017, doi: 10.1016/j.knosys.2017.06.033.

[40] K. Krysiak-Baltyn, T. Nordahl Petersen, K. Audouze, N. Jørgensen, L. Ängquist, and S. Brunak, "Compass: A hybrid method for clinical and biobank data mining," *J. Biomed. Inform.*, vol. 47, pp. 160–170, 2014, doi: https://doi.org/10.1016/j.jbi.2013.10.007.

[41] C. Sevilla-Salcedo, V. Imani, P. M. Olmos, V. Gómez-Verdejo, and J. Tohka, "Multi-task longitudinal forecasting with missing values on Alzheimer's disease," *Comput. Methods Programs Biomed.*, vol. 226, p. 107056, 2022, doi: https://doi.org/10.1016/j.cmpb.2022.107056.

[42] Z. Zhao, Z. Wu, Y. Zheng, and P. Ma, "Recurrent neural networks for atmospheric noise removal from InSAR time series with missing values," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, no. August, pp. 227–237, 2021, doi: 10.1016/j.isprsjprs.2021.08.009.

[43] T. Levy-Loboda, M. Rav-Acha, A. Katz, and N. Nissim, "Cardio-ML: Detection of malicious clinical programmings aimed at cardiac implantable electronic devices based on machine learning and a missing values resemblance framework," *Artif. Intell. Med.*, vol. 122, p. 102200, 2021, doi: https://doi.org/10.1016/j.artmed.2021.102200.

[44] N. Basurto, C. Cambra, and Á. Herrero, "Improving the detection of robot anomalies by handling data irregularities," *Neurocomputing*, vol. 459, pp. 419–431, 2021, doi: https://doi.org/10.1016/j.neucom.2020.05.101.

[45] X. Lai, X. Wu, and L. Zhang, "Autoencoder-based multi-task learning for imputation and classification of incomplete data," *Appl. Soft Comput.*, vol. 98, p. 106838, 2021, doi: https://doi.org/10.1016/j.asoc.2020.106838.

[46] S. Piri, "Missing care: A framework to address the issue of frequent missing values;The case of a clinical decision support system for Parkinson's disease," *Decis. Support Syst.*, vol. 136, p. 113339, 2020, doi: https://doi.org/10.1016/j.dss.2020.113339.

[47] D. Xu, P. J.-H. Hu, T.-S. Huang, X. Fang, and C.-C. Hsu, "A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management," *J. Biomed. Inform.*, vol. 111, p. 103576, 2020, doi: https://doi.org/10.1016/j.jbi.2020.103576.

[48] A. Janghorbani and M. H. Moradi, "Fuzzy Evidential Network and Its Application

as Medical Prognosis and Diagnosis Models," *J. Biomed. Inform.*, vol. 72, pp. 96–107, 2017, doi: https://doi.org/10.1016/j.jbi.2017.07.004.

[49] Y. Li and A. Ngom, "Classification approach based on non-negative least squares," *Neurocomputing*, vol. 118, pp. 41–57, 2013, doi: 10.1016/j.neucom.2013.02.012.

[50] J. Abawajy, A. Kelarev, M. Chowdhury, A. Stranieri, and H. F. Jelinek, "Predicting cardiac autonomic neuropathy category for diabetic data with missing values," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1328–1333, 2013, doi: https://doi.org/10.1016/j.compbiomed.2013.07.002.

[51] Y. Qu, J. Rong, Z. Li, and K. Chen, "ST-A-PGCL: Spatiotemporal adaptive periodical graph contrastive learning for traffic prediction under real scenarios," *Knowledge-Based Syst.*, vol. 272, p. 110591, 2023, doi: https://doi.org/10.1016/j.knosys.2023.110591.

[52] R. Rios *et al.*, "Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: Insights from REFINE SPECT registry," *Comput. Biol. Med.*, vol. 145, p. 105449, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105449.

[53] F. Shah, A. Castelltort, and A. Laurent, "Handling missing values for mining gradual patterns from NoSQL graph databases," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 523–538, 2020, doi: https://doi.org/10.1016/j.future.2019.10.004.

[54] G. Choi, H.-S. Oh, and D. Kim, "Enhancement of variational mode decomposition with missing values," *Signal Processing*, vol. 142, pp. 75–86, 2018, doi: https://doi.org/10.1016/j.sigpro.2017.07.007.

[55] M. L. Yadav and B. Roychoudhury, "Handling missing values: A study of popular imputation packages in R," *Knowledge-Based Syst.*, vol. 160, no. June, pp. 104–118, 2018, doi: 10.1016/j.knosys.2018.06.012.

[56] J. Xia *et al.*, "Adjusted weight voting algorithm for random forests in handling missing values," *Pattern Recognit.*, vol. 69, pp. 52–60, 2017, doi: 10.1016/j.patcog.2017.04.005.

[57] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016, doi: 10.1016/j.knosys.2015.12.006.

[58] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing*, vol. 205, pp. 152–164, 2016, doi: 10.1016/j.neucom.2016.04.015.

[59] M. L. Yadav and B. Roychoudhury, "Handling missing values: A study of popular imputation packages in R," *Knowledge-Based Syst.*, vol. 160, no. December 2017, pp. 104–118, 2018, doi: 10.1016/j.knosys.2018.06.012.

[60] R. Abolghasemi, R. Khadka, P. G. Lind, P. Engelstad, E. H. Viedma, and A. Yazidi, "Predicting missing pairwise preferences from similarity features in group decision making," *Knowledge-Based Syst.*, vol. 256, p. 109860, 2022, doi: 10.1016/j.knosys.2022.109860.