Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare

Amelia Ritahani Ismail [1]*, Nadzurah Zainal Abidin [2], Mhd Khaled Maen [3]
[1, 2] Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, P.O Box 10, 50728 Kuala Lumpur, Malaysia.
[3] Department of Information Technology, ITC, hus 1, 2 och 4, Lägerhyddsvägen 2, Box 337, 751 05 UPPSALA
Email: [1] amelia@iium.edu.my, [2] nadzurah.zabidin@gmail.com, [3] mhdkhaled.maen.0643@student.uu.se
*Corresponding Author

Abstract—**Missing data is one of the most common issues encountered in data cleaning process especially when dealing with medical dataset. A real collected dataset is prone to be incomplete, inconsistent, noisy and redundant due to potential reasons such as human errors, instrumental failures, and adverse death. Therefore, to accurately deal with incomplete data, a sophisticated algorithm is proposed to impute those missing values. Many machine learning algorithms have been applied to impute missing data with plausible values. However, among all machine learning imputation algorithms, KNN algorithm has been widely adopted as an imputation for missing data due to its robustness and simplicity and it is also a promising method to outperform other machine learning methods. This paper provides a comprehensive review of different imputation techniques used to replace the missing data. The goal of the review paper is to bring specific attention to potential improvements to existing methods and provide readers with a better grasps of imputation technique trends.**

Keywords—Review; Missing Data; Imputation; Machine Learning; Healthcare

I. Introduction

Prior to data mining process, data cleaning is an essential process to improve efficiency of analyzing data and to ensure the quality. One of the major tasks in data cleaning phase is to impute missing data. Data cleaning is a process of detecting and removing errors and inconsistencies from data in order to improve the quality of data [1], [2]. Most healthcare datasets were found to be incomplete, which double suffers to perform task of medical data mining. This is due to the fact that incorrect prediction measures may leads to improper medical treatment [3]–[5]. As reported by Yelipe and other author, there are seven research issues when handling with healthcare datasets, which are imputation of missing values, dimensionality, elimination of outliers, handling imbalanced datasets, attribute reduction, choice of classification approaches, and elimination of outliers [6]–[8]. The Figure 1 below shows the seven research problems when dealing with medical datasets.

Imputation of missing data is a mandatory step since any analysis of data cannot perform with incomplete dataset. Ignoring the step may results to invalid conclusions. Missing values contribute in imposing undesirable outcome, especially when it leads to biased estimations [10]–[12]. In data mining, imputation is a process of replacing missing data with plausible values.

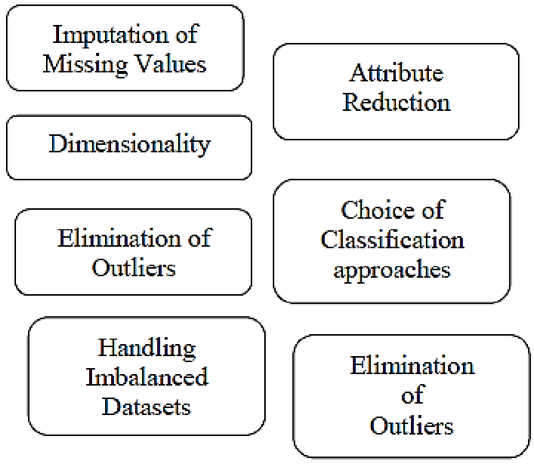

Fig. 1. Research problems when dealing with medical datasets [9]

Although the imputation techniques for missing data has been debatable for decades, there are small number of studies examining on the quality of the most proposed machine learning algorithms to impute missing data. There are several practices to deal and address missing data, and techniques of imputation missing values can be discovered. One of the practices that this paper attempts to discuss is an imputation techniques through machine learning algorithms [13]–[15]. A proper method of imputing can help to improve the quality of datasets for analyzing better healthcare decision.

The purpose of this paper is to analyze the depth of techniques used to impute missing data with the aid of machine learning classifiers. Besides, this paper attempts to review on the techniques available when dealing with missing values with imputation. Specifically, the research contribution for this paper are to review articles that deals with missing data in healthcare domains and learn the imputation techniques trends.

Next section will be summarizing the related work and number of research from primary studies on imputing missing data with machine learning classifiers. The rest of the paper is organized as follows; section 3 on the methodology used to conduct the SLR; section 4 is describing the taxonomy of primary studies collected based on query; section 5 and 6 attempt to discuss on the results gathered from the primary studies.

II. RELATED WORK

Missing values are the most common problem in all field area, be no exception to healthcare. In healthcare, the presence of missing values can be challenging issues especially in supporting healthcare decision [16]. The controversy of imputation has been discussed since 1998, however, the evolution of imputation with machine learning rise after a while especially in healthcare industry domain [17]–[20]. As discussed, imputation of missing data can be discovered through statistical and machine learning, and both carry strength and limitations to deal with it. However, evidence shows that statistical techniques of imputing still show bias in estimates missing values and far suffer from loss of information [21]–[24].

Numerous studies presented statistical methods to impute missing data such as multiple imputation, mean imputation and expected maximization [25]–[29]. Nevertheless, these methods will not preserve any relationship or association between variables in a dataset [30], [31]. Although imputation with statistical methods is allowable, it is strongly recommended to use any alternative approach that provide more accurate parameter estimates.

Healthcare has produced up to 60 percent of missing values, which may inflict to an outcast analysis result and real-world decision making. The researchers view this issues seriously, where missing values commonly makes the knowledge discovery a very difficult task [32], [33]. In the paper, the authors presented a model for an imputation for any type of missing data using three different algorithms; Amelia, FURIA, and MICE. Based on the experiment between three algorithms, MICE perform better to impute real healthcare dataset [34]–[36].

In addition, a researcher suggested new concept in imputing missing values which requires to cluster all medical records without missing values. A rationale reason beyond this new approach was imputation will be performed more accurate if all similar medical reports were in one cluster [9]. Apart from that, the authors highlighted the importance of imputation without merely eliminates the missing values in medical records.

An author [37] also suggested three imputation classification techniques to assess the performance of machine learning classifiers with missing values using Bagged Tree imputation (BTI). The adoption of the Bagged Tree imputation approach resulted in the highest accuracy for all three supervised classifiers such as neural network, random forest (RF), and support vector machine (SVM). The paper reported that RF has the greatest performance, followed by neural network and SVM [38], [39].

Several significant contributions have been made by [2], [40] focused on adopting data mining techniques for imputing the missing values. Among three proposed techniques such as random forest, decision tree and linear regression; the investigation resulted random forest outperforms decision tree and linear regression [41].

There are many studies discussed on the comparison between machine learning classifiers. These researchers usually analyze which among proposed algorithms performs best in imputing missing values. Authors acknowledged the efficiency of machine learning algorithms to impute missing values for different domains namely; multilayer perceptron (MLP), self-organizing map (SOM), decision tree (DT), K-nearest neighbors (KNN), FURIA, support vector machine (SVM), and K-means [42]. However, among the algorithms mentioned, the author investigates five classifiers particularly decision tree (DT), KNN, SVM, FURIA, and K-means to compare the performance with the traditional statistical methods. The results were compared with the most commonly used statistical approach to handle missing values mean-mode imputation. In the paper, several approaches proposed for imputation with machine learning outperforms other traditional statistical imputation methods in regard to the sensitivity and accuracy [18], [43].

Another accepted article is a journal entitled "A Comparison of Six Methods for Missing Data Imputation" [44]. The journal analyzed the performance of six machine learning classifiers namely, Mean, K-nearest neighbor (KNN), Fuzzy K-means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (bPCA), multiple imputations by chained equations (MICE). This paper demonstrates the imputation approach using four real medical datasets such as iris, E. coli, breast cancer1, and breast cancer2.

Another researchers investigate a set of machine learning imputation technique in particular naïve bayes, SVM, artificial neural network (ANN), KNN, decision trees, MLP, and k-means clustering [45]–[51]. The objectives of the paper was to assess the prediction of proposed techniques for top deadliest diseases [52]. All the machine learning algorithms mentioned were evaluated using three criteria namely; accuracy, sensitivity and specificity. The result shows that a higher evaluation parameter gives a better prediction and performance for kidney dialysis.

III. RESEARCH METHOD

A Systematic Literature Review (SLR) means of identifying, evaluating and interpreting relevant studies to a particular question or specific field [53]. The review process of this SLR follow closely to a platform named Parsifal (https://parsif.al/). Parsifal is a tool to assists researchers on conducting three crucial phases of SLR such as review planning, conducting, and documenting a report. Parsifal is helpful in terms of aiding the review process followed Kitchenham and Charters's procedure [54].

A. Research Question

RQ1: What evidence are there on imputing missing data techniques using machine learning algorithms for healthcare domain?
RQ2: Which machine learning algorithms are effective to optimize and improve for imputing missing data?
RQ3: How effective machine learning algorithms in imputing missing data?

The research questions were formulated with the aid of an approach called PICOC (population, interventions, comparison, outcomes, and context).

TABLE I. PICOC TABLE

PICOC	SCOPE
Population	Review of imputation technique for healthcare domain
Intervention	Imputation technique through machine learning classifiers
Comparisons	Quality of proposed machine learning classifiers in imputing missing data
Outcomes	Optimization of most frequently proposed machine learning classifiers
Context	Machine learning imputation techniques

B. Source Selection

The subject covered in this systematic literature review is healthcare. The main steps in selecting relevant source to review this paper include screening and filtering. Initially, each retrieved paper will be screened out if the articles were unrelated and does not provide sufficient information regarding imputation technique with machine learning. In the second iteration, the remaining paper were filtered by removing the duplicates articles, which left with only 536 articles. Third iteration, all the remaining paper were filtered by reading the article's title and abstract, and irrelevant studies were removed based on the inclusion and exclusion criteria mentioned as follows. The final iteration were to refer and read the full text articles and carefully reviewed. The following figure 2, were presenting on the flowchart to select the evidences.

A proper selection of the inclusion and exclusion criteria were derived to minimize the concerns of imputation approach to only machine learning algorithms. This is due to the fact that traditional statistical approach of imputation still produce bias in prediction measures.

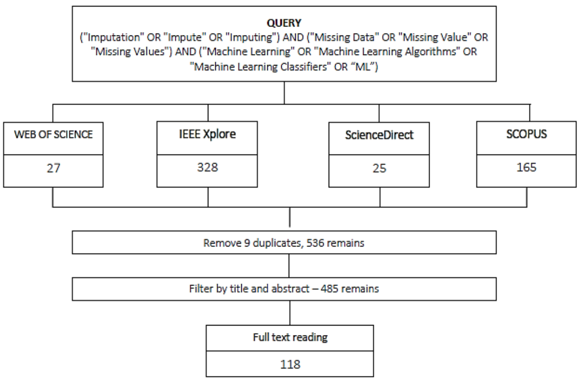

Fig. 2. Flowchart of Source Selection

Inclusion Criteria:

I1: The paper that experimenting on the improvement and optimization of proposed machine learning classifiers,

I2: Paper that published from 2011 to December 2020,

I3: The paper that describe imputation of missing values from only healthcare areas, and

I4: Paper published at any journal or conference paper.

Exclusion Criteria:

E1: The paper that published on imputing missing values using statistics approaches or techniques,

E2: The paper that do not described and written in English, and

E3: The proposed solution are new algorithms which composed from their respective domain tools.

C. Search String

The search string, which is expressed as a conjunction of three parts, was used to search within keywords, title, abstract and full text of a publication:

("Imputation" OR "Impute" OR "Imputing") AND ("Missing Data" OR "Missing Value" OR "Missing Values") AND ("Machine Learning" OR "Machine Learning Algorithms" OR "Machine Learning Classifiers" OR "ML").

D. Search Strategy

In order to identify the relevant studies, the key search terms detailed in the research questions and PICOC to search database. The major indexing databases are Scopus, IEEE Xplore, and Web of Science as table 2 below.

TABLE II. ONLINE DATABASES

Resource Name	Number of Studies
Web of Science	27
IEEE Xplore	328
ScienceDirect	25
Scopus	165
TOTAL	545

The total number of evidences from the first iteration captured from all domain areas, which includes healthcare. However, after analyzing all papers and organize the evidences into healthcare clusters only, we found out that only 118 are relevant to be shortlisted.

The table above aims to summarize and cluster all primary studies in order to build a comprehensive taxonomy. Taxonomy helps in transforming all 118 evidences into an organized manner and discovered that all the retrieved dataset can be classified into three groups; performance analysis of algorithms in imputation, improvement and optimization of imputation techniques, proposed on new methods for imputation with machine learning and review articles. A further finding is that 60 papers out of 118 was discussing on performance analysis between algorithms, 51 papers on optimization or improvement of imputation algorithms, 6 papers for proposing new solution with machine learning and 1 review paper. All summarization of retrieved papers can be found in the taxonomy discussed in the next section. This paper that explicitly describe on the taxonomy below will be further discussed in the next section.

IV. TAXONOMY ANALYSIS

To further analyze the literature this section provides taxonomy, challenges and motivations for missing data imputation with machine learning techniques as described in Figure 3. This is relevant in finding the gaps of the research that have been done. The literatures are further investigated imputation techniques based on:

- Enhancement work done in machine learning algorithms: Any literature on improving or optimizing for better performance accuracy, which will be conversed in Section 4.1.
- Proposals of new methods or framework: Section 4.2 will be discussing any suggestion made by authors as a new imputation algorithm.
- Performance analysis: Identify the common performance matrices to analyze machine learning algorithms performances, which will be discussed in Section 4.3.

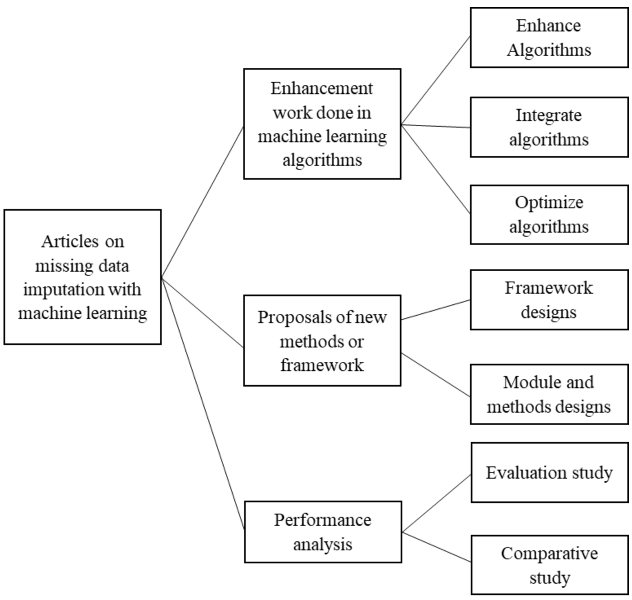

Fig. 3. Taxonomy of primary studies

A. *Enhancement work done in machine learning algorithms*

This category describe works that attempt to improve imputation performance by all means, either enhance existing algorithms, propose an integrated algorithm, ensemble the algorithms or optimize imputation algorithm with an optimization algorithm.

1) *Enhancing machine learning algorithms:*

The works fall under this category modified and expanded existing machine learning algorithm to enhance the performance of an algorithm to impute missing data. Some proposed methods were enhanced accordingly to suit with the experiment that authors are going to conduct. Authors [55]–[61] claimed that the enhancing machine learning algorithm process provide a better result than the existing algorithms. The algorithms that were proposed to be enhanced as imputation method are local least squares, fuzzy, and K-nearest neighbor. [42], [62]–[65] proposed a modify approach to estimate the missing values by combining the good features found in local least squares (LLS) imputation method. Another enhancing algorithm using LLS method were based on clustering techniques, and named after CLLS impute [63]. However, imputing missing value with LLS approach will only lead to iteratively adjusting found solution [66].

Saha et.al (2016) proposed a modification to the existing imputation named as Collaborative Filtering Based on Rough Set Theory (CFBRST) which uses fuzzy clustering technology to estimates missing values [67].

Several evidences had addressed to enhance the performance of traditional KNN with different approaches; using mutual information (MI) [68]–[71] and bagging methods [72]. However, both proposed algorithms insufficiently explored by experimenting using different weighting approaches and other machine learning methods for handling missing data.

2) *Integrating machine learning algorithms:*

These works most likely to observe the imputation techniques by augmenting two or more generic machine learning algorithms. At this circumstance, integrating machine learning algorithms were believe could obtain a better imputation performance than what could from any of the constituent machine learning algorithms [73].

Tran et.al proposed a combination of multiple imputation and ensemble learning to build an ensemble of classifiers for incomplete data classification tasks [74]. While [75] intended to ensemble classifiers with multiple imputation based on random subspace. Both suggested solutions achieves significantly better classification accuracy and perform quite well with large rate of missing values although there are inconsistent results for the imputed values [76], [77].

Nonetheless, two prior works [78] and [79] had done an experiment towards ensembles multiple imputation approaches with AdaBoost and bootstrapping respectively. Both suggested solutions perform slightly better than single imputation (mean, median, and KNN imputation) with only small percent of missingness ratio.

3) *Optimizing imputation algorithm approach machine learning algorithms:*

This category represents works which intended to improve the performance of imputation technique with the aid of optimization algorithm [80]–[86].

[80] had demonstrated a resemblances idea with this thesis by optimizing the K-nearest neighbors with optimization algorithm. The authors proposed a genetic algorithm (GA) to optimize KNN algorithm. The paper addressed on the usage of genetic optimization algorithm to KNN and investigated the impact of the accuracy of prediction missing data. The paper conducted a thorough analysis on the proposed algorithm and were further compared with the state-of-art method.

While, Kamiura et.al (2005) generally claim that adopting self-organizing maps (SOM) and GA algorithm may overcome an issue of missing item values and redundant data.

While, Priya et.al (2014) suggested an approach for optimizing the SVM imputation algorithm using principal component analysis (PCA). One author has developed a flexible and efficient algorithm to fill in the missing entries from the observed matrix using matrix completion approach [87].

Another famous optimization algorithm that were borrowed as an imputation algorithm is particle swarm optimization (PSO). Many novel approaches were conducted using different state-of-art algorithm such as decision tree [88], fuzzy c-means [89], and Bayesian network [90].

To sum up this section, the existing research has many problems in representing an extensive comparison with other similar machine learning methods and optimization algorithm for handling missing data.

B. Proposals of new methods or framework for imputation

In general, the works represents in this category proposed completely new approach of imputation claimed to estimate more accurately. Selected studies fall under two subcategories: framework designs and modules of methods designs.

1) Framework Designs:

Prior studies emphasized on proposing new framework designs for the purpose of assessing the reliability of specific prediction techniques [5], [73]–[76]. Some authors highlighted the main objective intending new approach to conform to the real-world problem of healthcare conditions.

Previous work presents a new framework design by integrating with other clustering algorithms to overcome the limitation of imputation techniques and samples of datasets [91]–[96]. The proposed imputation techniques claimed able to select appropriate subsets of the most relevant samples for better results of imputation value. Plus, the articles argued that the methods improve the accuracy in imputing missing values.

A recent study also has explored the issues with missing values and develop a new imputation method to maximize the accuracy in predicting. This new proposed solution were integrated with the best combination to maximize the discrimination margin of missing values [97].

2) Modules of Methods Designs:

Several prior articles had reported a new modules of methods that believe to estimate accurately and provide a better solution to complete microarray missing data [98]–[101].

There is also one study that suggest a new models which combine with other imputation modelling to make the process very flexible and robust [102]–[104]. By a simulation study and a real data analysis, the proposed model improves the imputation of missing data and uncertainty prediction estimation.

C. Performance Analysis

Performance analysis is a process of empirically evaluate algorithms to measure a success performance. The majority of prior research has emphasized on performance of machine learning algorithms to impute missing data. This section is divided into two (2) category: evaluation and comparative study.

1) Evaluation Study:

These research work mostly evaluate the performance of missing data imputation algorithms. Mainly, all the machine learning algorithms or proposed solution were evaluated using three useful parameters such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE) [105]–[108]. Generally, these parameters helps in evaluating the performance of predicting methods and to measure forecast accuracy [109], [110]. However, many of these scheme introduced to evaluate these algorithms are limited to measure the similarity between actual and imputed data. [111], [112] proposed to measure the success of imputed data from the perspective of normalized root mean square error and classification accuracy [113], [114] Authors also have driven the further development of imputation technique and evaluate the accuracy with either recall, accuracy, precision, F1 score, or receiver operations characteristic (ROC) [115]–[117]. These parameters are powerful in demonstrating and interpreting in order to measure the performance of imputing techniques with machine learning [118]–[120]. Despite, these evaluation parameters mentioned were not supported by an empirical analysis and hypothesis test for missing data imputation.

2) Comparative Study:

This work compared the performance of one or many proposed machine learning algorithms which outperforms other approaches and aims better results in imputation. Some of the author examined and compared the strength and limitations of other solution of imputation with their proposed solution algorithms [38], [121]–[125]. Besides, in short, a comparative study towards imputation can be classified into three categories (1) nature of datasets, (2) percentage of missing values, and (3) machine learning algorithms. Early studies have also suggested that the comparative study of an imputation approach should be based on the nature of the datasets [30], [126]. Nature of datasets can be referred to the scale of datasets and nature of missingness [127], [128]. Many articles agreed to the famous discussions by Rubin (1976) regarding the mechanisms of missingness. The mechanisms proposed by Rubin (1976) and colleagues [129] is highly referred which drawn conclusion that it is also highly influenced the selection of imputation methods. Three missingness mechanisms are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [130]–[133]. Another fundamental concept to consider is the classification of missing data towards the missingness mechanism [134]. Rubin (1976) is the first to introduce on three missing data mechanism in particular, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR can be described when the missingness is independent of the observed and missing values that is unrelated to the values of any variables [135]. MAR means there is a systematic relationship between the propensity of missing values and the observed data, the missingness is conditionally independent of the missing responses. In this case, missingness does not depend on the variable of interest, but it could depend on other observed variables. MNAR describe a situation where the propensity on both observed and missing data is dependent on the variable of interest. MNAR corresponds to all cases that neither MCAR nor MAR [136], [137]. An author even conclude that nature of dataset have a greater impact on the performance of imputation compared to the imputation method itself [30].

Accordingly, there are works focused on other factor such as percentage of missing values in order to compare the performance of imputation between algorithms. A comparative study was conducted to examine the association between performance of imputation and percentage of missing values [111], [138]. An experiment was conducted repeatedly over different rates of missing value from the range of 0% up to 90%. Theories proved that percentage of missing values strongly influence the performance of imputation [23]. A high percentage of missing values most likely to reduce the speed in imputing and imposed the imputation techniques. Three prior studies [109], [139] and [111] have experimented an imputation performance with huge differences of missing value rates, 58-85%, 5-60% and 0-20% respectively. Unlike [140] attempts to impute randomly missing values with small different rates of missing values, 1% to 5%. These rates are slightly small to observe the influences towards performance of missing values. Nonetheless, [141] use the actual percentage of missing values to evaluate the imputation performance.

V. Results and Discussion

All data retrieved using the query were demonstrated via a graphical presentation as shown in figure 4. The list of relevant studies in the bar chart as figure below shows that detailed out the number of evidences published by years.

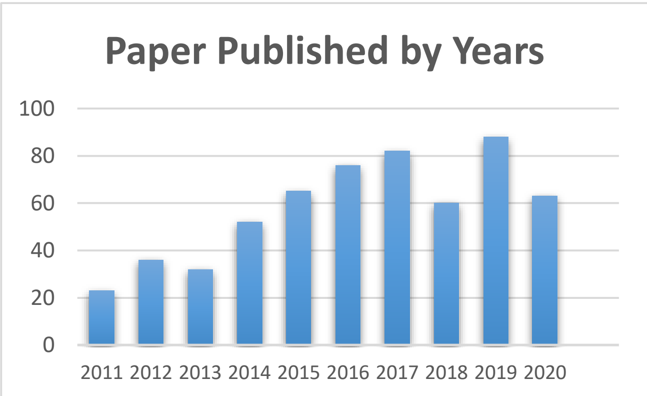

Fig. 4. Number of articles by years of publication

The figure above provides an information on the articles that were published in regard to imputation techniques with machine learning algorithms. The distribution of all evidences was collected for 10 years, starting from 2011 to July 2020. To date, the highest number of papers published on the topic were in 2019. Despite that, the numbers are presumed to growth in 2021. Among the ten years of publication, the least paper reported was on 2011, with only 23 papers published.

A systematic review is a research study that collects and looks at multiple studies. This SLR reviewed 118 evidences on imputation techniques through machine learning for healthcare domains. Most evidence compared their proposed solutions outperforms any other traditional machine learning approach, where indeed, there is no imputation techniques consistently outperforms every other. To conclude, the performance of imputation techniques with machine learning may be influenced by the nature of dataset instead of the techniques itself.

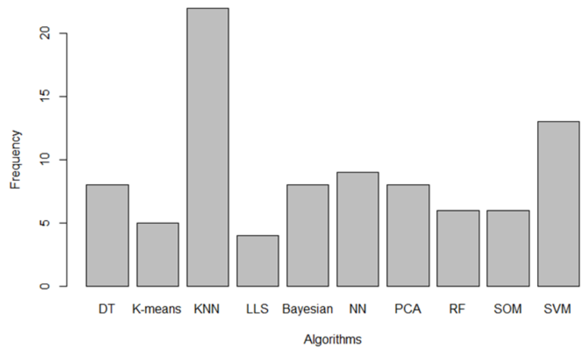

Fig. 5. Frequency of Algorithms for Healthcare domains

A comparative study on imputation were also experimented with various machine learning algorithms. The Figure 5 above illustrates the top 10 most used machine learning imputation algorithm among 44 algorithms in healthcare domain. The implications from these findings (figure 4 and 5) shows that KNN were the most frequent algorithm used to impute missing values. The fact of this matter is KNN claimed to be able to impute with any type and scale of a database. An advantage of KNN for an imputation routine is it will go through the entire healthcare dataset regardless the size of datasets. As the abbreviate meaning of KNN, it will find and replace the missing values on the basis of it nearest neighbors. The efficiency of this algorithms can be seen as it only requires to impute missing value captured by its related neighbors over its entire records [142]. Besides, in many cases, KNN algorithm outperform the other imputation methods namely support vector machine, naïve bayes, decision tree, self-organizing maps (SOM) and many more [74], [143]–[148]. The second most standout algorithm employed as imputation algorithm is Support Vector Machine (SVM). Authors [3], [46] described SVM imputation algorithm as produces fast, more accurate and robust classification results, however, [80] claimed that some approaches such as SVM, single value decomposition (SVD), and principal component analysis (PCA) are not compatible and causing negative effect on data with missing values. While [149] discussed on Bayesian limitations which appears as improper option in terms of accuracy and sensitive to imputation values. Decision tree and random forest were said to be shown its demerits in the sense of space limitations and low imputation accuracy if the size of a segment is small [150].

VI. Conclusion

Missing data is a universal problem in many research areas and may influence to the biased estimations and wrong conclusions. To overcome the drawbacks it produced, a process call ‘missing data imputation’ should be taken before proceeding to the next phase such in data mining. Besides, prior to data mining process, data cleaning is an essential process to improve efficiency of analyzing data and to ensure the quality. One of the major tasks in data cleaning phase is to impute missing data. Data cleaning is a process of detecting and removing errors and inconsistencies from data in order to improve the quality of data. Most healthcare datasets were found to be incomplete, which double suffers to perform task of medical data mining. This is due to the fact that incorrect prediction measures may leads to improper medical treatment.

A series of studies have been proposed machine learning as an imputation algorithm, and yet, there is no imputation algorithm that consistently outperforms others in every situation. However, selecting the most appropriate algorithm may significantly improve the accuracy of imputation results. Among all machine learning imputation algorithms, KNN algorithm has been widely adopted as an imputation for missing data and it is also a promising method to outperform other machine learning methods. KNN is a straightforward, yet powerful classification algorithm that computes a value estimates from the closest neighbors which has relatively high accuracy. The favorable points on KNN are simplicity, comprehensibility and scalability. However, despite the simplicity associated with KNN algorithm, several studies have well acknowledged that KNN suffers from high computational cost, greater storage requirements, and sensitivity to noise.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education for Fundamental Research Grant Scheme (FRGS): FRGS/1/2018/ICT02/UIAM/02/1.

REFERENCES

[1] E. Rahm and Hong Hai Do, "Data Cleaning : Problems and Current Approaches," IEEE Data Eng. Bull., no. January 2000, pp. 1–11, 2000.

[2] M. R. Vinutha and J. Chandrika, "Imputation as a technique for enhancing the quality of medical data," Int. J. Curr. Res. Rev., vol. 13, no. 5, pp. 91–95, 2021.

[3] M. Al Khaldy and C. Kambhampati, "Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset," Proc. SAI Intell. Syst. Conf. 2016.

[4] Y. Usharani and P. Sammulal, "A novel approach for imputation of missing values for mining medical datasets," 2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2015, 2016.

[5] Y. Zhu and X. Duan, "Predictive nursing helps improve treatment efficacy, treatment compliance, and quality of life in unstable angina pectoris patients," Am. J. Transl. Res., vol. 13, no. 4, pp. 3473–3479, 2021.

[6] A. Wang, H. Lim, S. Y. Cheng, and L. Xie, "ANTENNA, a Multi-Rank, Multi-Layered Recommender System for Inferring Reliable Drug-Gene-Disease Associations: Repurposing Diazoxide as a Targeted Anti-Cancer Therapy," IEEE/ACM Trans. Comput. Biol. Bioinforma., 2018.

[7] N. A. B. Kamisan, M. H. Lee, A. G. Hussin, and Y. Z. Zubairi, "Imputation techniques for incomplete load data based on seasonality and orientation of the missing values," Sains Malaysiana, vol. 49, no. 5, pp. 1165–1174, 2020.

[8] U. R. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," Comput. Electr. Eng., vol. 66, pp. 487–504, 2018.

[9] Y. Usharani and P. Sammulal, "A novel approach for imputation of missing values for mining medical datasets," 2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2015, 2016.

[10] N. Z. Zainal Abidin, A. R. Ismail, and N. A. Emran, "Performance Analysis of Machine Learning Algorithms for Missing Value Imputation," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, 2018.

[11] N. A. M. Pauzi, Y. B. Wah, S. M. Deni, S. K. N. A. Rahim, and Suhartono, "Comparison of single and mice imputation methods for missing values: A simulation study," Pertanika J. Sci. Technol., vol. 29, no. 2, pp. 979–998, 2021.

[12] L. Bargelloni et al., "Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream," Aquac. Reports, vol. 20, p. 100661, 2021.

[13] I. Erlyn and W. Rachmawan, "Optimization of Missing Value Imputation using Reinforcement Programming," Int. Electron. Symp. (IES), IEEE, pp. 128–133, 2015.

[14] C. Platias and G. Petasis, "A comparison of machine learning methods for data imputation," PervasiveHealth Pervasive Comput. Technol. Healthc., pp. 150–159, 2020.

[15] E. Thomas, T. and Rajabi, "A systematic review of machine learning-based missing value imputation techniques," Data Technol. Appl., vol. 55, no. 4, pp. 558–585, 2021.

[16] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," Health Qual. Life Outcomes, vol. 17, no. 1, pp. 1–9, 2019.

[17] S. Rostami, A. Kleszcz, D. Dimanov, and V. Katos, "A machine learning approach to dataset imputation for software vulnerabilities," Commun. Comput. Inf. Sci., vol. 1284 CCIS, no. September, pp. 25–36, 2020.

[18] A. W. Lo, K. W. Siah, and C. H. Wong, "Machine Learning with Statistical Imputation for Predicting Drug Approval," Harvard Data Sci. Rev., no. 1, 2019.

[19] A. Bhattacharjee and M. S. Bayzid, "Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices," BMC Genomics, vol. 21, no. 1, pp. 1–14, 2020.

[20] Y. C. Su, C. Y. Wu, C. H. Yang, B. S. Li, S. H. Moi, and Y. Da Lin, "Machine learning data imputation and prediction of foraging group size in a Kleptoparasitic spider," Mathematics, vol. 9, no. 4, pp. 1–16, 2021.

[21] N. Solomon, Y. Lokhnygina, and S. Halabi, "Comparison of regression imputation methods of baseline covariates that predict survival outcomes," J. Clin. Transl. Sci., vol. 5, no. 1, 2021.

[22] M. Khayati, A. Lerner, Z. Tymchenko, and P. Cudre´Mauroux, "Mind the gap: An experimental evaluation of imputation of missing values techniques in time series," Proc. VLDB Endow., vol. 13, no. 5, pp. 768–782, 2020.

[23] L. Huang, C. Wang, and N. A. Rosenberg, "The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations," Am. J. Hum. Genet., vol. 85, no. 5, pp. 692–698, 2009.

[24] R. A. Hughes, J. Heron, J. A. C. Sterne, and K. Tilling, "Accounting for missing data in statistical analyses: Multiple imputation is not always the answer," Int. J. Epidemiol., vol. 48, no. 4, pp. 1294–1304, 2019.

[25] B. O. Petrazzini, H. Naya, F. Lopez-Bello, G. Vazquez, and L. Spangenberg, "Evaluation of different approaches for missing data imputation on features associated to genomic data," BioData Min., vol. 14, no. 1, pp. 1–13, 2021.

[26] S. Siafis et al., "Imputing the number of responders from the mean and standard deviation of CGI-improvement in clinical trials investigating medications for autism spectrum disorder," Brain Sci., vol. 11, no. 7, 2021.

[27] J. B. Hardouin, R. Conroy, and V. Sébille, "Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data," BMC Med. Res. Methodol., vol. 11, pp. 1–13, 2011.

[28] L. Malan, C. M. Smuts, J. Baumgartner, and C. Ricci, "Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns," Nutr. Res., vol. 75, pp. 67–76, 2020.

[29] H. M. K. Ghomrawi, L. A. Mandl, J. Rutledge, M. M. Alexiades, and M. Mazumdar, "Is there a role for expectation maximization imputation in addressing missing data in research using WOMAC questionnaire? Comparison to the Standard mean approach and a tutorial," BMC Musculoskelet. Disord., vol. 12, no. 1, p. 109, 2011.

[30] Y. Liu and V. Gopalakrishnan, "An overview and evaluation of recent machine learning imputation methods using cardiac imaging data," Data, vol. 2, no. 1, 2017.

[31] S. Javadi, A. Bahrampour, M. M. Saber, B. Garrusi, and M. R. Baneshi, "Evaluation of Four Multiple Imputation Methods for Handling Missing Binary Outcome Data in the Presence of an Interaction between a Dummy and a Continuous Variable," J. Probab. Stat., vol. 2021, pp. 1–14, 2021.

[32] S. I. Chowdhury, M. H., Islam, M. K., & Khan, "Imputation of Missing Healthcare Data," Comput. Inf. Technol. (ICCIT), 2017 20th Int. Conf., pp. 1–6, 2017.

[33] L. A. Kahale et al., "Potential impact of missing outcome data on treatment effects in systematic reviews: Imputation study," BMJ, vol. 370, pp. 1–10, 2020.

[34] T. Köse, S. Özgür, E. Coşgun, A. Keskinoğlu, P. Keskinoğlu, and D. Mrozek, "Effect of Missing Data Imputation on Deep Learning Prediction Performance for Vesicoureteral Reflux and Recurrent Urinary Tract Infection Clinical Study," Biomed Res. Int., vol. 2020, 2020.

[35] R. Wijesuriya, M. Moreno-Betancur, J. B. Carlin, and K. J. Lee, "Evaluation of approaches for multiple imputation of three-level data," BMC Med. Res. Methodol., vol. 20, no. 1, pp. 1–15, 2020.

[36] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, "MICE vs PPCA: Missing data imputation in healthcare," Informatics Med. Unlocked, vol. 17, no. November, p. 100275, 2019.

[37] I. Jordanov, N. Petrov, and A. Petrozziello, "Classifiers Accuracy Improvement Based on Missing Data Imputation," J. Artif. Intell. Soft Comput. Res., vol. 8, no. 1, pp. 31–48, 2018.

[38] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva, "Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study," BMC Bioinformatics, vol. 20, no. 1, pp. 1–11, 2019.

[39] T. Etoeharnowo and M. H. J. A. Van Os, "A Random Forest Approach for Dealing with Missingness: a Case Study in Primary Care Data," Leidin Inst. Adv. Comput. Sci., 2020.

[40] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," BMC Med. Res. Methodol., vol. 20, no. 1, pp. 1–12, 2020.

[41] S. M. Mostafa, "Imputing missing values using cumulative linear regression," CAAI Trans. Intell. Technol., vol. 4, no. 3, pp. 182–200, 2019.

[42] T. R. Sivapriya, "Imputation And Classification Of Missing Data Using Least Square Support Vector Machines – A New Approach In Dementia Diagnosis," IJARAI - Int. J. Adv. Res. Artif. Intell., vol. 1, no. 4, pp. 29–34, 2012.

[43] M. Rahman and D. N. Davis, "Machine Learning Based Missing Value Imputation Method for Clinical Datasets," IAENG Trans. Eng. Technol. Springer Netherlands, vol. 247, no. January, 2013.

[44] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," J. Biom. Biostat., vol. 6, no. 1, pp. 1–6, 2015.

[45] W. Seffens et al., "Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study," Bioinform. Biol. Insights, vol. 9s3, pp. 43–54, 2015.

[46] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, "Fast Imbalanced Classification of Healthcare Data with Missing Values," pp. 1–13, 2015.

[47] M. W. Huang, W. C. Lin, and C. F. Tsai, "Outlier Removal in Model-Based Missing Value Imputation for Medical Datasets," J. Healthc. Eng., vol. 2018, 2018.

[48] J. Bektaş, T. Ibrikçi, and İ. T. Özcan, "The impact of imputation procedures with machine learning methods on the performance of classifiers: An application to coronary artery disease data including missing values," Biomed. Res., vol. 29, no. 13, pp. 2780–2785, 2018.

[49] C. Cheng and H. Huang, "A Distance-Threshold kNN Method for Imputing Medical Data Missing Values," vol. 7, no. 1, pp. 13–17, 2019.

[50] S. F. Huang and C. H. Cheng, "A safe-region imputation method for handling medical data with missing values," Symmetry MDPI, vol. 12, no. 11, pp. 1–19, 2020.

[51] T. Rockel, D. W. Joenssen, and U. Bankhofer, "Decision Trees for the Imputation of Categorical Data," Kit Sci. Publ. , vol. 2, no. 1, pp. 1–15, 2017.

[52] A. Ameta and K. Jain, "Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review," Int. J. Eng. Comput. Sci., vol. 6, no. 2, pp. 20376–20378, 2017.

[53] S. Nair, J. L. De La Vara, M. Sabetzadeh, and L. Briand, "An extended systematic literature review on provision of evidence for safety certification," Inf. Softw. Technol., vol. 56, no. 7, pp. 689–717, 2014.

[54] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Engineering, vol. 2, p. 1051, 2007.

[55] S. Bose, C. Das, T. Gangopadhyay, and S. Chattopadhyay, "A modified local least squares-based missing value estimation method in microarray gene expression data," Int. Conf. Adv. Comput. Netw. Secur., pp. 18–23, 2013.

[56] Y. Y. Choi, H. Shon, Y. J. Byon, D. K. Kim, and S. Kang, "Enhanced application of principal component analysis in machine learning for imputation of missing traffic data," Appl. Sci., vol. 9, no. 10, pp. 1–15, 2019.

[57] X. Su, R. Greiner, T. M. Khoshgoftaar, and A. Napolitano, "Using classifier-based nominal imputation to improve machine learning," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6634 LNAI, no. PART 1, pp. 124–135, 2011.

[58] F. D. Atem, E. Sampene, and T. J. Greene, "Improved conditional imputation for linear regression with a randomly censored predictor," Stat. Methods Med. Res., vol. 28, no. 2, pp. 432–444, 2019.

[59] A. Sundararajan and A. I. Sarwat, "Evaluation of Missing Data Imputation Methods for an Enhanced Distributed PV Generation Prediction," Proc. Futur. Technol. Conf. 2019, pp. 590–609, 2AD.

[60] M. Alber et al., "Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences," npj Digit. Med., vol. 2, no. 1, 2019.

[61] M. G. Signorini, N. Pini, A. Malovini, R. Bellazzi, and G. Magenes, "Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring," Comput. Methods Programs Biomed., vol. 185, 2020.

[62] S. Nikfalazar, C. H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," Knowl. Inf. Syst., vol. 62, no. 6, pp. 2419–2437, 2020.

[63] P. Keerin and W. Kurutach, "An Improvement of Missing Value Imputation in DNA Microarray Data Using Cluster-based LLS Method," 2013 13th Int. Symp. Commun. Inf. Technol., pp. 559–564, 2013.

[64] M. F. Dzulkalnine and R. Sallehuddin, "Missing data imputation with fuzzy feature selection for diabetes dataset," SN Appl. Sci., vol. 1, no. 4, pp. 1–12, 2019.

[65] L. E. Chai et al., "Investigating the effects of imputation methods for modelling gene networks using a dynamic Bayesian network from gene expression data," Malaysian J. Med. Sci., vol. 21, no. 2, pp. 20–27, 2014.

[66] I. Wasito and B. Mirkin, "Nearest neighbours in least-squares data imputation algorithms with different missing patterns," Comput. Stat. Data Anal., vol. 50, no. 4, pp. 926–949, 2006.

[67] S. Saha, S. Bandopadhyay, A. Ghosh, and K. N. Dey, "An improved fuzzy based approach to impute missing values in DNA microarray gene expression data with collaborative filtering," 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016, pp. 911–916, 2016.

[68] Pedro J. Garcia-Laencina, A. R. F. Vidal, and J.-L. Sancho-Gomez, "A Robust Approach For Classifying Unknown Data in Medical Diagnosis Problems," in 2008 World Automation Congress, 2008.

[69] E. Tavazzi, S. Daberdaku, R. Vasta, A. Calvo, A. Chiò, and B. Di Camillo, "Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data with an adaptive k-nearest neighbours approach," BMC Med. Inform. Decis. Mak., vol. 20, no.

Suppl 5, pp. 1–23, 2020.

[70] P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, “K-nearest neighbours based on mutual information for incomplete data classification,” ESANN 2008 Proceedings, 16th Eur. Symp. Artif. Neural Networks - Adv. Comput. Intell. Learn., no. May 2014, pp. 37–42, 2008.

[71] P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, “K nearest neighbours with mutual information for simultaneous classification and missing data imputation,” Neurocomputing, vol. 72, no. 7–9, pp. 1483–1493, 2009.

[72] V. Kumutha and S. Palaniammal, “An Enhanced Approach on Handling Missing Values Using Bagging k-NN Imputation,” 2013 Int. Conf. Comput. Commun. Informatics, pp. 1–8, 2013.

[73] K. Mehrabani-Zeinabad, M. Doostfatemeh, and S. M. T. Ayatollahi, “An efficient and effective model to handle missing data in classification,” Biomed Res. Int., vol. 2020, 2020.

[74] C. T. Tran, M. Zhang, P. Andreae, and B. Xue, “Multiple Imputation and Ensemble Learning for Classification with Incomplete Data,” Intell. Evol. Syst., vol. 187, no. December 2017, 2009.

[75] L. Nanni, A. Lumini, and S. Brahnam, “A classifier ensemble approach for the missing feature problem,” Artif. Intell. Med., vol. 55, no. 1, pp. 37–50, 2012.

[76] D. An, R. J. A. Little, and J. W. McNally, “A multiple imputation approach to disclosure limitation for high-age individuals in longitudinal studies,” Stat. Med., vol. 29, no. 17, pp. 1769–1778, 2010.

[77] C. G. Schuetz, “Using neuroimaging to predict relapse to smoking: role of possible moderators and mediators.,” Int. J. Methods Psychiatr. Res., vol. 17 Suppl 1, no. 1, pp. S78–S82, 2008.

[78] B. Conroy, L. Eshelman, C. Potes, and M. Xu-Wilson, “A dynamic ensemble approach to robust classification in the presence of missing data,” Mach. Learn., vol. 102, no. 3, pp. 443–463, 2016.

[79] S. S. Khan, A. Ahmad, and A. Mihailidis, “Bootstrapping and Multiple Imputation Ensemble Approaches for Missing Data,” Cornell Univ., no. Mi, pp. 1–17, 2016.

[80] H. De Silva and A. S. Perera, “Missing Data Imputation using Evolutionary k-Nearest Neighbor Algorithm for Gene Expression Data,” pp. 141–146, 2016.

[81] N. Kamiura, A. Ohtsuka, H. Tanii, T. Isokawa, and N. Matsui, “On Detection of Hematopoietic Tumors Using Self Organizing Maps and Genetic Algorithms,” in 2005 IEEE International Conference on Systems, Man and Cybernetics, 2005.

[82] M. Priya and P. R. Kumar, “Intelligent Approaches for Prognosticating Atherosclerotic and Non-Atherosclerotic Individuals,” 2014 Int. Conf. Commun. Signal Process., pp. 691–695, 2014.

[83] P. Almasinejad, A. Golabpour, M. Reza, M. Meybodi, K. Mirzaie, and A. Khosravi, “A Dynamic Model for Imputing Missing Medical Data : A Multiobjective Particle Swarm Optimization Algorithm,” J. Healthc. Eng., vol. 2021, no. i, 2021.

[84] X. Wang, W. Li, Y. Sun, S. Milanovic, M. Kon, and J. E. Castrillon-Candas, “Multilevel Stochastic Optimization for Imputation in Massive Medical Data Records,” arXiv, 2021.

[85] A. R. Ismail, N. A. Aziz, A. M. Ralib, N. Z. Abidin, and S. S. Bashath, “A particle swarm optimization levy flight algorithm for imputation of missing creatinine dataset,” Int. J. Adv. Intell. Informatics, vol. 7, no. 2, pp. 225–236, 2021.

[86] J. W. and C. L. Anderson, Deborah K., Liang, “An imputation-regularized optimization algorithm for high dimensional missing data problems and beyond,” Physiol. Behav., vol. 176, no. 5, pp. 139–148, 2017.

[87] C. Ke et al., “Prognostics of surgical site infections using dynamic health data,” J. Biomed. Inform., vol. 65, pp. 22–33, 2017.

[88] V. S. H. Rao, S. Member, and M. N. Kumar, “Novel Approaches for Predicting Risk Factors of Atherosclerosis,” IEEE J. Biomed. Heal. Informatics, vol. 17, no. 1, pp. 183–189, 2013.

[89] M. Najib and N. A. Samat, “FCMPSO : An Imputation for Missing Data Features in Heart Disease Classification,” IOP Conf. Ser. Mater. Sci. Eng., 2017.

[90] A. Nekouie and M. H. Moattar, “Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization,” J. King Saud Univ. - Comput. Inf. Sci., 2018.

[91] M. Suresh, R. Taib, Y. Zhao, and W. Jin, “Sharpening the BLADE: Missing Data Imputation Using Supervised Machine Learning,” Data Integr. Partnersh. Aust., no. July, pp. 215–227, 2019.

[92] S. I. Khan and A. S. M. L. Hoque, “SICE: an improved missing data imputation technique,” J. Big Data, vol. 7, no. 1, 2020.

[93] Y. Guo, Z. Liu, P. Krishnswamy, and S. Ramasamy, “Bayesian Recurrent Framework for Missing Data Imputation and Prediction with Clinical Time Series,” arXiv, 2019.

[94] K. Skivington et al., “A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance,” BMJ Open, no. 2018, p. n2061, 2021.

[95] Q. Zhao and D. Lee, “HICCUP : Hierarchical Clustering Based Value Imputation using Heterogeneous Gene Expression Microarray Datasets,” 2007.

[96] D. Bertsimas, A. Orfanoudaki, and C. Pawlowski, Imputation of clinical covariates in time series, vol. 110, no. 1. Springer US, 2020.

[97] R. Kumar, T. Chen, M. Hardt, D. Beymer, K. Brannon, and T. Syeda-Mahmood, “Multiple Kernel Completion and its application to cardiac disease discrimination,” Proc. - Int. Symp. Biomed. Imaging, pp. 764–767, 2013.

[98] T. Le, T. Altman, and K. J. Gardiner, “Probability-based Imputation Method for Fuzzy Cluster Analysis of Gene Expression Microarray Data,” 2012 Ninth Int. Conf. Inf. Technol. - New Gener., pp. 42–47, 2012.

[99] M. Liu, Y. Gao, P. T. Yap, and D. Shen, “Multi-Hypergraph Learning for Incomplete Multimodality Data,” IEEE J. Biomed. Heal. Informatics, vol. 22, no. 4, pp. 1197–1208, 2018.

[100] P. Valarmathie and K. Dinakaran, “An efficient technique for missing value imputation in microarray gene expression data,” no. Icccs 114, pp. 073–080, 2014.

[101] X. Wu, H. Akbarzadeh Khorshidi, U. Aickelin, Z. Edib, and M. Peate, “Imputation techniques on missing values in breast cancer treatment and fertility data,” Heal. Inf. Sci. Syst., vol. 7, no. 1, pp. 1–8, 2019.

[102] M. O. Prates, “Spatial extreme learning machines: An application on prediction of disease counts,” Stat. Methods Med. Res., p. 96228021876798, 2018.

[103] L. Jin et al., “A comparative study of evaluating missing value imputation methods in label-free proteomics,” Sci. Rep., vol. 11, no. 1, pp. 1–11, 2021.

[104] M. Zitnik et al., “Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities,” HHS Public Access, no. 50, pp. 71–91, 2020.

[105] A. Ghandeharioun et al., “Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data,” 2017 Seventh Int. Conf. Affect. Comput. Intell. Interact., pp. 325–332, 2017.

[106] C. Velasco-Gallego and I. Lazakis, “Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study,” Ocean Eng., vol. 218, no. July, p. 108261, 2020.

[107] M. A. H., N. D. Nur, N. Md Tahir, Z. Iffah Abd Latiff, M. Huzaimy Jusoh, and Y. Akimasa, “Missing data imputation of MAGDAS-9’s ground electromagnetism with supervised machine learning and conventional statistical analysis models,” Alexandria Eng. J., vol. 61, no. 1, pp. 937–947, 2022.

[108] O. A. Alade, R. Sallehuddin, and A. Selamat, “Empirical Performance Evaluation of Imputation Techniques using Medical Dataset,” IOP Conf. Ser. Mater. Sci. Eng., vol. 551, no. 1, pp. 0–5, 2019.

[109] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman, “Techniques to cope with missing data in host-pathogen protein interaction prediction,” Bioinformatics, vol. 28, no. 18, pp. 466–472, 2012.

[110] E. Kontopantelis, I. R. White, M. Sperrin, and I. Buchan, “Outcome-sensitive multiple imputation: A simulation study,” BMC Med. Res. Methodol., vol. 17, no. 1, pp. 1–13, 2017.

[111] V. F. Ghoneim, N. H. Solouma, and Y. M. Kadah, “Evaluation of missing values imputation methods in cDNA microarrays based on classification accuracy,” 2011 1st Middle East Conf. Biomed. Eng. MECBME 2011, pp. 367–370, 2011.

[112] C. Chang, Y. Deng, X. Jiang, and Q. Long, “Multiple imputation for analysis of incomplete data in distributed health data networks,” Nat. Commun., vol. 11, no. 1, pp. 1–11, 2020.

[113] V. F. Ghoneim, N. H. Solouma, and Y. M. Kadah, “The Impact of

Missing Values Imputation Methods in cDNA Microarrays on Downstream Data Analysis," in 28th NATIONAL RADIO SCIENCE CONFERENCE 28th NATIONAL RADIO SCIENCE CONFERENCE, 2011, no. Nrsc.

[114] C. Y. Guo, Y. C. Yang, and Y. H. Chen, "The Optimal Machine Learning-Based Missing Data Imputation for the Cox Proportional Hazard Model," Front. Public Heal., vol. 9, no. July, pp. 1–8, 2021.

[115] A. Colubri, T. Silver, T. Fradet, K. Retzepi, B. Fry, and P. Sabeti, "Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients," PLoS Negl. Trop. Dis., vol. 10, no. 3, pp. 1–17, 2016.

[116] M. Mitra and R. K. Samanta, "A Study on UCI Hepatitis Disease Dataset Using Soft Computing," AMSE Journals IIETA, vol. 78, pp. 467–477, 2017.

[117] M.-K. Suh, J. Woodbridge, M. Lan, A. Bui, L. S. Evangelista, and M. Sarrafzadeh, "Missing Data Imputation for Remote CHF Patient Monitoring Systems," Conf. Proc. IEEE Eng. Med. Biol. Soc. NIH Public Access, pp. 3184–3187, 2012.

[118] D. M. Hondula et al., "A respiratory alert model for the Shenandoah Valley, Virginia, USA," Int. J. Biometeorol., vol. 57, no. 1, pp. 91–105, 2013.

[119] 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.

[120] S. Phung, A. Kumar, and J. Kim, "A deep learning technique for imputing missing healthcare data," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 6513–6516, 2019.

[121] J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," Artif. Intell. Med., vol. 50, no. 2, pp. 105–115, 2010.

[122] K. Grace-Martin, "Limitations of Common Solutions to Missing Data," Cornell Statistical Consulting Unit, no. November, 2001.

[123] J. A. Saunders, N. Morrow-howell, E. Spitznagel, P. Doré, E. K. Proctor, and R. Pescarino, "Imputing Missing Data: A Comparison of Methods for Social Work Researchers," in Social Work Research, 2006, pp. 19–31.

[124] M. H. Huque, J. B. Carlin, J. A. Simpson, and K. J. Lee, "A comparison of multiple imputation methods for missing data in longitudinal studies 01 Mathematical Sciences," BMC Med. Res. Methodol., vol. 18, no. 1, pp. 1–16, 2018.

[125] A. K. Waljee et al., "Comparison of imputation methods for missing laboratory data in medicine," BMJ Open, vol. 3, no. 8, pp. 1–8, 2013.

[126] M. Soley-bori, "Dealing with missing data: Key assumptions and methods for applied analysis," PM931 Dir. Study Heal. Policy Manag., no. 4, p. 20, 2013.

[127] S. Haji-maghsoudi, A. Haghdoost, A. Rastegari, and M. R. Baneshi, "Influence of Pattern of Missing Data on Performance of Imputation Methods : An Example Using National Data on Drug Injection in Prisons," Int. J. Heal. Policy Manag., vol. 1, no. 1, pp. 69–77, 2013.

[128] S. N. Payrovnaziri, A. Xing, S. Salman, X. Liu, J. Bian, and Z. He, "The Impact of Imputation on the Interpretations of Prediction Models: A Case Study on Mortality Prediction for Patients with Acute Myocardial Infarction," AMIA ... Annu. Symp. proceedings. AMIA Symp., vol. 2021, pp. 465–474, 2021.

[129] D. B. Rubin and R. J. A. Little, Statistical Analysis with Missing Data, Second Edi. New York: A John Wiley & Sons, Inc., Publication, 2002.

[130] H. Kang, "The prevention and handling of the missing data," Korean J. Anesthesiol., vol. 64, no. 5, pp. 402–406, 2013.

[131] B. Leurent, M. Gomes, S. Cro, N. Wiles, and J. R. Carpenter, "Reference-based multiple imputation for missing data sensitivity analyses in trial-based cost-effectiveness analysis," Heal. Econ. (United Kingdom), vol. 29, no. 2, pp. 171–184, 2020.

[132] A. B. Pedersen et al., "Missing data and multiple imputation in clinical epidemiological research," Clin. Epidemiol., vol. 9, pp. 157–166, 2017.

[133] M. R.-R. Jacques-Emmanuel Galimard, Sylvie Chevret, Camelia Protopopescu, "A multiple imputation approach for MNAR mechanisms compatible with Heckman's model," Wiley Online Libr., 2016.

[134] M. Pampaka, G. Hutcheson, and J. Williams, "Handling missing data : analysis of a challenging data set using multiple imputation," Int. J. Res. Method Educ., vol. 7288, no. 39:1, pp. 19–37, 2016.

[135] M. N. N. Ramli, A. S. Yahaya, N. A. Ramli, N. F. F. . Yusof, and M. M. A. Abdullah, "Roles of Imputation Methods for Filling the Missing Values : A Review," in International Conference of Advanced Materials Engineering and Technology (ICAMET 2013), 2013, no. November.

[136] M. G. Kenward, "The handling of missing data in clinical trials," Clin. Investig. (Lond)., vol. 3, no. 3, pp. 241–250, 2013.

[137] D. C. Howell, "The Treatment of Missing Data," pp. 1–44, 2000.

[138] M. Mera-Gaona, U. Neumann, R. Vargas-Canas, and D. M. López, "Evaluating the impact of multivariate imputation by MICE in feature selection," PLoS One, vol. 16, no. 7 July, pp. 1–28, 2021.

[139] G. Wang, Z. Deng, and K.-S. Choi, "Tackling missing data in community health studies using additive LS-SVM classifier," IEEE J. Biomed. Heal. Informatics, vol. 22, no. 2, pp. 1–1, 2016.

[140] M. S. B. Sehgal, I. Gondal, and L. Dooley, "K-Ranked Covariance Based Missing Values Estimation for Microarray Data Classification," in Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 2001, pp. 8–13.

[141] M. S. B. Sehgal, I. Gondal, and L. S. Dooley, "Gene expression Collateral missing value imputation : a new robust missing value estimation algorithm for microarray data," Bioinformatics, vol. 21, no. 10, pp. 2417–2423, 2005.

[142] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms : a critical evaluation," BMC Med. Inform. Decis. Mak., vol. 16, no. Suppl 3, 2016.

[143] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," in Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), 2004, no. 1995, pp. 1–9.

[144] W. C. W. Chen et al., "Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data," J. Biomed. Informatics, vol. 33, no. 2, pp. 557–573, 2016.

[145] J. Huang et al., "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study," J. Syst. Softw., vol. 132, pp. 226–252, 2017.

[146] R. Samant and S. Rao, "Effects of Missing Data Imputation on Classifier Accuracy," Int. J. Eng. Res. Technol., vol. 2, no. 11, pp. 264–266, 2013.

[147] M. G. Rahman and M. Z. Islam, "A decision tree-based missing value imputation technique for data pre-processing," Conf. Res. Pract. Inf. Technol. Ser., vol. 121, pp. 41–50, 2010.

[148] Per Jönsson and Claes Wohlin, "An Evaluation of k -Nearest Neighbour Imputation Using Likert Data," 10th Int. Symp. Softw. Metrics, 2004. Proc., 2004.

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

[150] 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.