https://doi.org/10.48047/AFJBS.6.Si3.2024.3130-3150

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Missing Value Estimation Methods for Classification of Arrhythmia using Deep Learning: Review Study

Biswajit Brahma[1], Ricky Mohanty[2], Subrat Kumar Parida[3], Avijit Mondal[4]

[1]McKesson Corporation,California
[2]School of Information System, ASBM University, Bhubaneswar, India
[3]Department of Computer Science and Engineering,Centurion University of Technology and Management,
Vizianagaram, Andhra Pradesh,India
[4]School of Information System, ASBM University, Bhubaneswar, India

Article Info
Volume 6, Issue Si3, 2024
Received: 30 May 2024
Accepted : 20 June 2024
Doi:
10.48047/AFJBS.6.Si3.2024.3130-3150

Abstract

The increasing prevalence of cardiovascular diseases (CVDs) has become a major health concern. Arrhythmia is the deadliest heart condition of all cardiovascular disorders. Thus, timely and precise arrhythmia diagnosis is critical in preventing heart disease and abrupt cardiac death. Arrhythmia can be discovered on an electrocardiogram (ECG) by observing irregular heart electrical activity. The heart's electrical activity is recorded as an ECG signal, which contains both normal and pathological information. Classification of ECG patterns is critical for automatically diagnosing cardiac illness. This paper discusses the various learning approaches for automatically distinguishing different types of heartbeats. According to reported studies, the convolutional neural network (CNN) model is the best option for classifying arrhythmia. An ensemble of depth wise separable convolutional (DSC) neural networks achieves the highest classification accuracy, 99.88%.

Keywords: Datasets,GA (Genetic algorithm), Feature selection, Information Gain, Missing Values Imputation, RMSE (Root mean square error)

1.**Introduction**

Missing Value Imputation (MVI), intended as the principal solution strategy for datasets having one or more missing attribute's values, has recently been the subject of a large number of studies. The addition of MVI improves the performance of Machine Learning (ML) models and calls for a thorough analysis of the MVI approaches used for various workloads and datasets. It will serve as a guide for beginners on how to create an efficient ML-based decision-making system for use in a variety of applications. In the literature that has been published in the last ten years, the state-of-the-art MVI approaches will be thoroughly reviewed and analysed in this article. The well-known Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology is used to choose 191 publications that were published for review. We include pertinent definitions for those articles in our summary.

2.Data Mining

Data mining or big data analysis is recognized as a crucial and difficult task for many applications in daily life, where a specific dataset for a preferred topic is gathered to conduct such an analysis. Any learnable decision-making system for automated classification or regression tasks must start with a dataset.But in order to create a general trained model, it is necessary to remove the practical dataset's unusually high proportion of missing values, irregular patterns (outliers), and redundancy with one or more variables. Tradition dictates that the former missing values be encoded as NaNs, blanks, undefined, null, or any other type of placeholder [1].Such missingness can be attributed to a wide range of sources in the datasets.However, there are three types of mechanisms for the missing values in the datasets [2], namely, Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). Let us consider an attribute variable Ii connected to the ith sample as in Eq. (1).

$$I_i = \begin{cases} 1 & if\ ith\ sample\ responds \\ 0, & otherwise \end{cases} [1]$$

Consider an attribute variable $I_i$connected to the i^{th} sample as in Eq. (2.1).In response modeling, we suppose that P $\{I_i = 1\} = \theta_i$. The response probability θ_i may rely on a set of known P auxiliary information x_i= (x_1 , …, x_{ji}, …, x_{pi}) or samples' true label yi.These include incorrect and inaccurate data entry, data availability issues, data gathering issues, missing features, missing files, incomplete information, etc. When an instance's or observation's likelihood of having a missing value for an attribute does not depend on either the known values or the lost data, as is the case when the missing values are randomly distributed across all observations, the first mechanism, known as MCAR, takes place [2].

The absence of data is unrelated to any study variable in MCAR. So, a representative sample of all participants is made up of the persons who have all of the observed data, given a specific intervention. Assume that the probability i is unrelated to the research variable yi. In that circumstance, regardless of the dataset, observed values, or unobserved values, the missing values come within the MCAR category. For instance, if a schedule was lost, another one may be substituted from among the filled schedules, chosen at random. The second process, known as MAR, takes place when an instance's likelihood of missing an attribute value may be influenced by that property [2].

In other words, the lost data on a partially missing variable (y_p) is not related to the values of yp itself, but rather to some other fully observed variables (x_w) in the analytic model. Assume that y_i's observable values but not its missing values determine the response probability i. In that situation, the missing data are referred to as MAR, depending on the seen data rather than the unobserved values. If, for instance, food consumption is missing from household surveys but the size of the household is present, we can still determine the missing food consumption by fitting a linear regression. The chance of an instance having a missing value for an attribute may vary on the value of that property [2], which results in the third remaining mechanism, known as NMAR. The data fall within the category of MNAR if they do not characterize with MCAR or MAR. The nonresponse is known as NMAR if idepend on the missing variable y_i, where the missingness is dependent on the observed values of the dataset and the missing value. For instance, individuals who earn a lot of money are less likely to disclose it, while those who are infected with the Human Immunodeficiency Virus (HIV) are less likely to disclose it. However, a more recent article in 2020, which can be found in [2], reviewed those methods in great detail. The primary focus of this paper does not include a detailed examination of these missingness processes. Regardless of the origin, dealing with missing data is critical since any statistical findings based on a dataset with non-random missing values may be skewed. Missing data also contributes an inherent degree of ambiguity into a statistical analysis. In addition, many Machine Learning (ML) methods do not tolerate missing value data [3].

This paper examines and analyses Missing Value Imputation (MVI) methodologies in depth, as well as their judgements. The technological concepts, with accompanying benefits and drawbacks, of numerous MVI schemes, as well as mathematical formulations of their evaluation metrics, are meticulously presented to assist researchers in gathering such materials in a single study. This page also covers many ML models and associated evaluation criteria, with descriptions and formulations that are suitably succinct. This review also includes information on current trends in MVI methodologies used, MVI evaluation, ML models, and ML model evaluation. This study also indicates how the successful application of appropriate MVI methods with ML models improves the performance of decision-making actions, providing the results from numerous papers with various MVI and ML models for various datasets in different domains. However, the core contributions of this review article on the MVI methods will be supportive for the research community for the following contexts:

- The community of researchers benefits greatly from the essential knowledge on missing values, their missingness mechanisms, and MVI methodologies with assessment metrics.
- Compiles twelve years' worth of missingness imputation data (from 2010 to 2021), including an assessment of the most recent trends in the use of MVI approaches.
- Various ML models are included, together with performance metrics from the chosen papers, demonstrating trends in the most often used models and evaluation criteria.
- Presents the findings from numerous studies pertaining to various methodologies and domain datasets to demonstrate that adding MVI methods can enhance the performance of ML models.
- Provides acceptable explanations of MVI methods and ML models, together with their evaluations and a number of suggestions for choosing them. This information may help researchers of all skill levels create comprehensive decision-making systems.

The literature search for this study is carried out using the well-known Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) technique [4].PRISMA is a minimum set of evidence-based components that authors can use to present different kinds of systematic reviews and meta-analyses. It enables transparent and complete study reporting and is generally used to estimate the issues with a healthcare intervention. Using the search "missing value imputation" on "Google Scholar," a total of n = 428 papers for the MVI method were discovered. A few duplications in the literature are evident in the initially detected articles (n = 57), which are then eliminated to leave n = 371 items. Publications authored in languages other than English were next subjected to the first screening phase (first-level exclusion) (n = 11) and articles with a review, database, or letter format (n=12) are excluded.In this cycle, only articles from the years 2010 through August 2021 are kept. The total number of articles produced by this exclusion is n = 266 (see Fig. 1). When all the selected papers from the earlier steps are scrutinised, it is discovered that n = 37 of them have irrelevant study objectives and n = 21 of them use different techniques. N = 208 articles are made available by this second exclusion (see Fig. 1). Finally, n = 17 papers are further removed because neither they nor the other applied tactics openly present the MVI methodologies. N = 191 articles are finally included to carry out this targeted review process after the ensuing identification, screening, and eligibility testing.

2.1Overview of Missing Value Imputation Techniques

This section applies the PRISMA approach to provide a complete overview and review of various MVI approaches (see Section 2.1.3) and their assessment metrics (see Section 2.2) from the chosen papers (see Fig. 1). As mentioned in the previous section, a total of 191 papers are chosen for the review process. These 191 articles, which are shown in Fig. 2 as the year-by-year publication numbers for the MVI techniques from 2010 to 2021, are exhibited. This data shows that, especially after 2017, the number of publications on the MVI method per year increases with a positive slope. As a result of ourdecision to choose articles through August 2021, it is remarkable to note that the number of publications in 2021 has decreased.

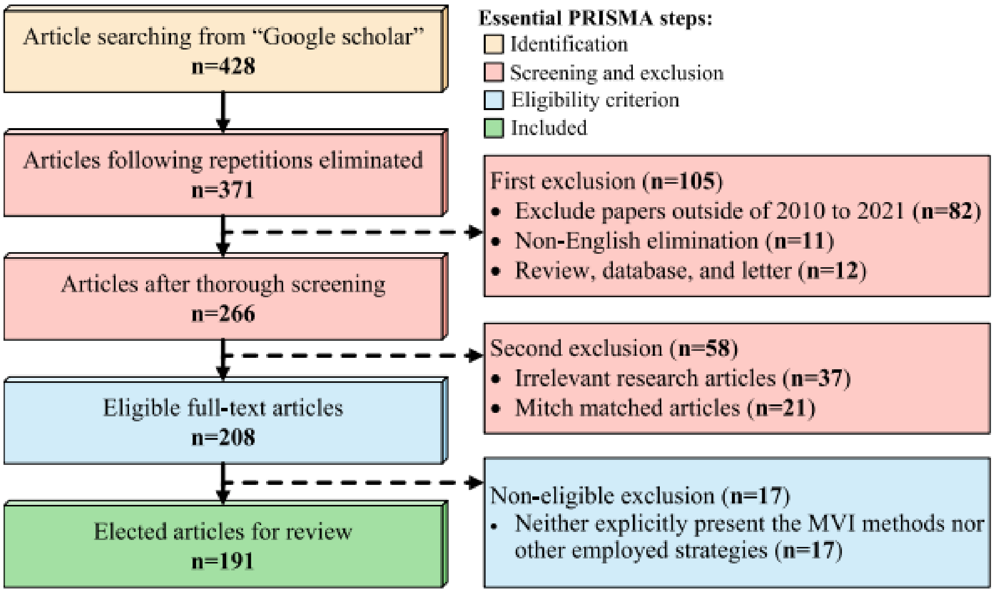

Article searching from "Google scholar"
n=428
Articles following repetitions eliminated
n=371
Articles after thorough screening
n=266
Eligible full-text articles
n=208
Elected articles for review
n=191
Essential PRISMA steps:
□ Identification
🟥 Screening and exclusion
▨ Eligibility criterion
🟩 Included
First exclusion (n=105)
• Exclude papers outside of 2010 to 2021 (n=82)
• Non-English elimination (n=11)
• Review, database, and letter (n=12)
Second exclusion (n=58)
• Irrelevant research articles (n=37)
• Mitch matched articles (n=21)
Non-eligible exclusion (n=17)
• Neither explicitly present the MVI methods nor other employed strategies (n=17)

Fig. 1. The employed PRISMA method for article searching procedure, where we demonstrate the evidence for the article's inclusion and exclusion criteria.

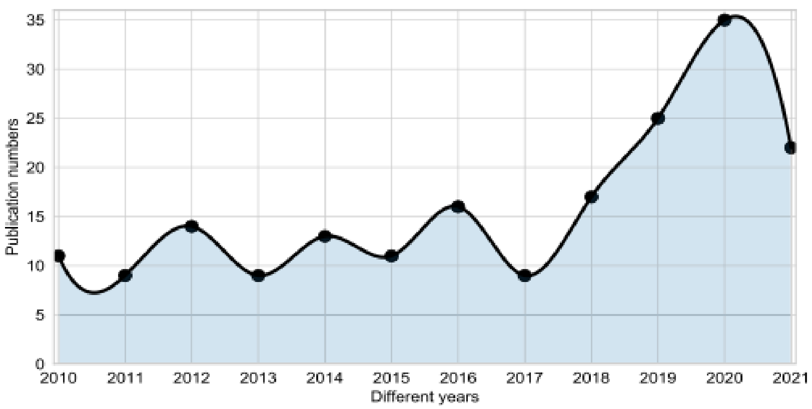


Fig. 2. Presentation of the number of per-year publications on the MVI method for the ML-based models. The number of publications in 2021 has been reduced due to the selection until August.

As a result, a thorough examination of the techniques of MVIs in those massive papers is required. Many writers experimented with imputation success by displaying missing value simulations over a given dataset at various missingness rates. Again, some researchers looked at low missing rates, such as less than 30%, while others looked at high missing rates, such as 5 80% [2]. This article only covers the MVIs' methodological perspectives in the selected 191 papers. Lin and Tsai [2], on the other hand, discussed and analysed various studies that accounted for distinct missing rates: less than 30%, between 30% and 50%, and greater than 50%. When the dataset contains a little amount of missingness (for example, less than 10% or 15% of the total dataset), the missing samples can be removed, a process known as case deletion, without materially impacting the final decision-making result. When the missing rate approaches 15%, however, special caution must be taken. However, not every domain dataset meets this condition, especially when little quantities of missing data transmit essential information, as in the case of records holding exceptionally large volumes of consumer spending data. In contrast to the case elimination strategy, MVI is the most commonly employed approach to solving the riddle of the fragmented dataset. MVI is a statistical or machine learning-based strategy for replacing missing values with new ones. Section 2.1 describes MVI methods and outlines the various MVI algorithms.

2.2 Imputation methods

The experimental block diagram of the general MVI technique, which divides each incomplete dataset into complete and missing sets, is shown in Figure 3. In order to replace the missing values in the missing dataset, one of the various MVI algorithms (shown in Fig. 4) uses the previously complete data to learn (train) parameters and estimate the right values. The simplest strategy for evaluating the imputation algorithms is to assess the discrepancies between the actual and imputed numbers. The alternative method involves classifying or grouping the outputted whole dataset and looking at the results' metrics. The following Section 2.4 examines the evaluation's specifics. A vast number of MVI approaches [5-45] are discovered by a review of the literature and are loosely classified into two [2]: statistical and ML-based MVI strategies, as shown in Fig. 4. These two types of MVI approaches are further subdivided into many algorithm types (shown in Fig. 4), which simplifies comparison presenting. Table 1 provides a general discussion of the approaches mentioned in Fig. 4, as well as whether they are supervised or unsupervised. The fundamental idea for applying those methods is presented in Table 1 along with any associated benefits or drawbacks. As shown in Table 1, the algorithmic information of those MVI methods for implementation is also connected by supplying the appropriate references or citations.

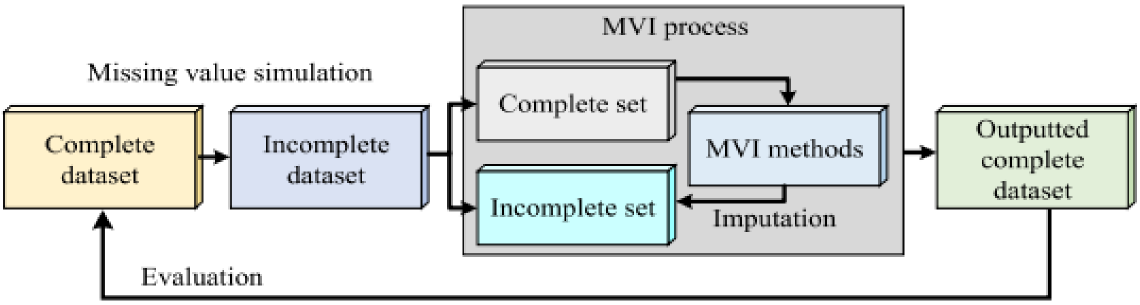


Fig. 3. The typical experimental configuration for MVI procedures to impute the missing values in any attributes [2].

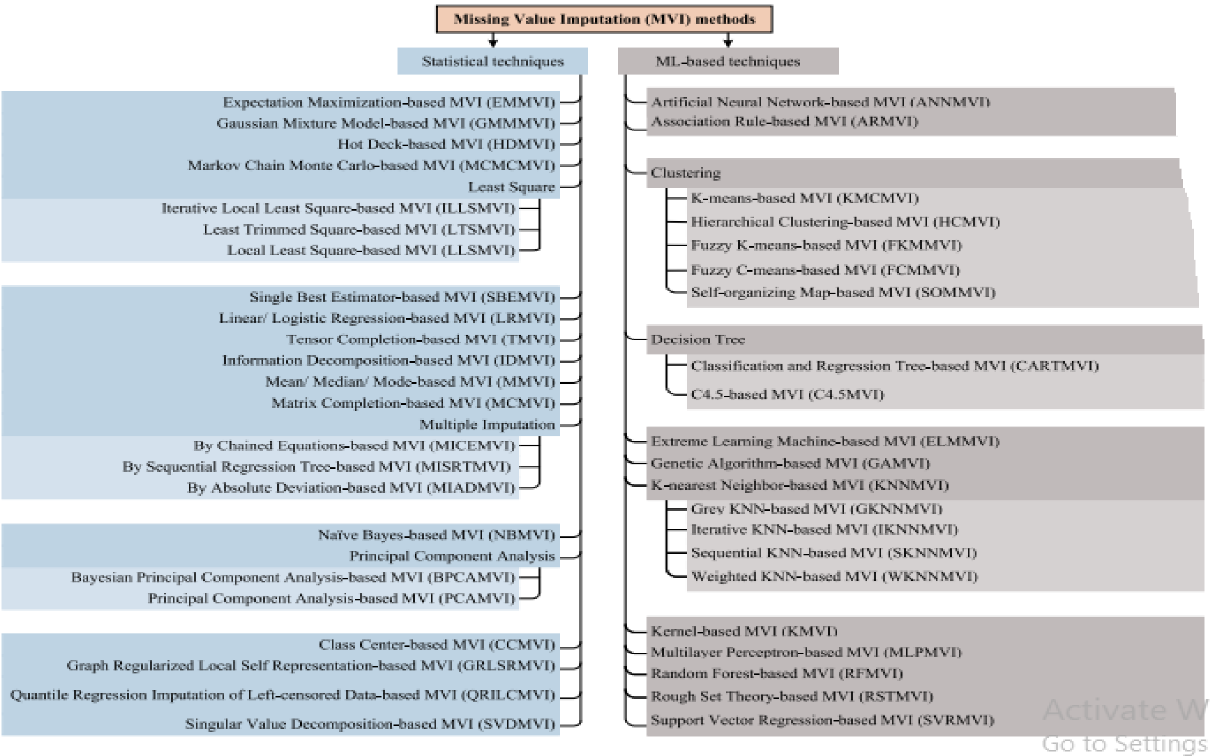

Missing Value Imputation (MVI) methods
Statistical techniques
ML-based techniques
Expectation Maximization-based MVI (EMMVI)
Gaussian Mixture Model-based MVI (GMMMVI)
Hot Deck-based MVI (HDMVI)
Markov Chain Monte Carlo-based MVI (MCMCMVI)
Least Square
Iterative Local Least Square-based MVI (ILLSMVI)
Least Trimmed Square-based MVI (LTSMVI)
Local Least Square-based MVI (LLSMVI)
Single Best Estimator-based MVI (SBEMVI)
Linear/ Logistic Regression-based MVI (LRMVI)
Tensor Completion-based MVI (TMVI)
Information Decomposition-based MVI (IDMVI)
Mean/ Median/ Mode-based MVI (MMVI)
Matrix Completion-based MVI (MCMVI)
Multiple Imputation
By Chained Equations-based MVI (MICEMVI)
By Sequential Regression Tree-based MVI (MISRTMVI)
By Absolute Deviation-based MVI (MIADMVI)
Naïve Bayes-based MVI (NBMVI)
Principal Component Analysis
Bayesian Principal Component Analysis-based MVI (BPCAMVI)
Principal Component Analysis-based MVI (PCAMVI)
Class Center-based MVI (CCMVI)
Graph Regularized Local Self Representation-based MVI (GRLSRMVI)
Quantile Regression Imputation of Left-censored Data-based MVI (QRILCMVI)
Singular Value Decomposition-based MVI (SVDMVI)
Artificial Neural Network-based MVI (ANNMVI)
Association Rule-based MVI (ARMVI)
Clustering
K-means-based MVI (KMCMVI)
Hierarchical Clustering-based MVI (HCMVI)
Fuzzy K-means-based MVI (FKMMVI)
Fuzzy C-means-based MVI (FCMMVI)
Self-organizing Map-based MVI (SOMMVI)
Decision Tree
Classification and Regression Tree-based MVI (CARTMVI)
C4.5-based MVI (C4.5MVI)
Extreme Learning Machine-based MVI (ELMMVI)
Genetic Algorithm-based MVI (GAMVI)
K-nearest Neighbor-based MVI (KNNMVI)
Grey KNN-based MVI (GKNNMVI)
Iterative KNN-based MVI (IKNNMVI)
Sequential KNN-based MVI (SKNNMVI)
Weighted KNN-based MVI (WKNNMVI)
Kernel-based MVI (KMVI)
Multilayer Perceptron-based MVI (MLPMVI)
Random Forest-based MVI (RFMVI)
Rough Set Theory-based MVI (RSTMVI)
Support Vector Regression-based MVI (SVRMVI)

Fig. 4. The categorized tree exhibition of the commonly employed MVI methods, available in the literature.

Table 1 lists all of the MVI methods shown in Figure 4 along with its citation, basic imputation mechanism, and specific advantages or downsides.

The unsupervised and supervised MVI approaches are denoted by the symbols cross-mark(✗) andcheck-mark(✓).

Methods (Supervised?)	Technical concept	Remarks
Statistic-based MVI methods		
EMMVI[5](✗)	Iterates between the E-step and M-step to estimate parameters directly, maximising the total data log-likelihood function.	Very sluggish convergence and inapplicable when the primary goals are confidence intervals of calculated parameters.
GMMMVI [6] (✗)	Determines the missing value as data generation iteratively from the cluster, satisfying the condition of the log-likelihood function.	Provides unsatisfactory imputed values if the class likelihood and its cluster have a heavily shared or common region between the classes
HDMVI [46] (✗)	Imputes missing values by adopting values from similar, but complete records of the same dataset, employing an object called the donor.	Although Gower's coefficient determines a donor, a single donor might be selected to provide multiple recipients, creating redundancy.
MCMCMVI [8] (✗)	Assigns the missing values from likelihood estimation based on Bayesian inference using the imputation and posterior steps.	Not straightforward to determine the convergence, but reliable to replace missing value either a low or high percentage of missing data.
ILLSMVI [9] (✓)	Predicts the missing data from the regressors, wherein each iteration, one variable is used as a response variable, and other variables serve as regressors.	Requires to be more careful in this setting to ensure that the separate regression models are consistent among them.
LTSMVI [10] (✓)	Starts from the KNN and is followed by iterative regressions in a transformed isometric log-ratio space.	Allows estimating the missing cells in the response with the multivariate information and outliers.
LLSMVI [11] (✓)	Least square-based method, applying many regressions using K neighbors taking K coherent genes instead of weighting or averaging K coherent genes.	Applicable to various biological and chemical data; robust and accurate missing value estimation method.
SBEMVI [12] (✗)	Simple substitution method, selecting neighboring proximity elements (min. distance from target place).	Neglecting variances, as it considers only neighboring elements, can produce erroneous imputation.
LRMVI [13] (✓)	Predicts the missing values from the regression coefficients of the fitted line or curve.	Massively dependent on the outliers present in the training samples.
TMVI [14] (✓)	Fills the missing elements, where the low rank of the tensor is often necessary to restrict the degree of freedom to avoid being an underdetermined problem.	More stable and accurate, especially when the missing sample entries are minimal, and it can propagate structure to fill more significant missing regions.
IDMVI [15] (✗)	Imputes values, combined with	Often needs KNN to find more

	multidimensional data with the help of the fuzzy membership function.	information when the fuzzy membership cannot transfer information.
MMVI [25] (✗)	Replaces values with a mean/ median/ mode value of non-missing elements in the corresponding variable.	Reduces the variance of the imputed variables and does not preserve correlations between variables.
MCMVI [16] (✗)	Fills the values by solving a Matrix Completion (MC) problem through inexact augmented Lagrange multipliers, considering MC as an optimization problem.	It is a hurdle to find stopping criteria during optimization due to the oscillation of convergence-minima for some kinds of datasets.
MICEMVI [17] (✗)	Assigns values from the regression model iteratively until reaching the stopping criterion, where missing positions are initially filled with mean and variance.	It has a significant advantage over other methods in terms of flexibility but does not have the same theoretical grounds as other methods.
MISRTMVI [18] (✗)	Combined Multiple Imputations (MI) and single tree prediction, decomposing the prediction error to the expected value and variance of the MI prediction.	Raises ambiguity when the imputation number increases and the correlation among prediction models decreases.
MIADMVI [19] (✗)	Uses outer covariates to fill missing values iteratively based on the fitted conditional model until meeting the ending criterion	. Demands slightly more data analytic capacity and skills of programming.
NBMVI [20] (✓)	Fills the missing values according to the Bayes theorem-based probability estimation from real data.	Imputation error value depends on the size of complete data, instances with no missing values.
BPCAMVI [21] (✗)	Uses probabilistic latent variables applying PCA regression, Bayesian estimation, and EM algorithms.	Not computationally efficient as it applies those three algorithms simultaneously.
PCAMVI [22] (✗)	Generates imputed data from a PCA model and are evaluated by projecting them to the principal axis.	Only higher variances are considered to minimize the errors; some crucial information may be lost.
CCMVI [23] (✗)	Each class center and its distances are marked to identify a threshold used for missing value imputation.	Requires very less imputation time than the ML-based methods
GRLSRMVI [24] (✗)	Rough-filled values are refined using local self-defined cost function by linearly mixing the nearby samples and graph regularization, forcing the nearby samples' deviation.	Provides a global and uniform framework for recovering missing values, incorporating local structure information of data.
ML-based MVI methods		
QRILCMVI [25] (✗)	Imputes missing values by randomly drawing from a truncated distribution estimated by quantile regression, carrying log-transformation to gain accuracy.	Designed explicitly for left-censored data.
SVDMVI [26] (✗)	Fills from a linear mixture of the k	Has the significant advantages of

	most eigen variables iteratively until a certain convergence threshold.	handling datasets of various sizes, mixed types, and faster.
ANNMVI [27] (✓)	The missing values are filled from the prediction of the end-to-end auto-encoder.	Separates overlapped clusters of data but are heavily affected by highly skewed data.
ARMVI [28] (✗)	Association rules describe the dependency links among data entries in a dataset where all data, including the missing ones, are filled according to that rules.	Only database entries which exactly match the candidate patterns may contribute to the support of the candidate pattern.
KMCMVI [29] (✗)	Iteratively tries to partition the dataset into K distinct clusters based on the sum of the squared distance between the data points and the cluster's centroid.	Can group more or less linearly separable clusters and do not work well with the global cluster.
HCMVI [30] (✗)	Build a hierarchy of clusters based on a greedy method, which falls into two types: agglomerative (bottom-up) and divisive (top-down) approaches.	No apriori information about the number of clusters required and easy to implement, requiring more time complexity (n 2 (n)) for n samples.
FKMMVI [31] (✗)	An approach for exploring the structure of a set of patterns, when the clusters are overlapping and are supposed to reflect the complex nature of data.	Makes the resulting algorithms less susceptible to getting stuck in a local minimum and provides a better tool when the clusters are not well separated.
FCMMVI [32] (✗)	Data assigns to every cluster with likelihood and repeats until the likelihood is maximized and converged.	Allows one datum to belong to two or more clusters and is frequently employed in pattern recognition
SOMMVI [33] (✗)	Allows one datum to two or more clusters and is frequently employed in pattern recognition.	Relies on a predefined distance in feature space and does not behave so gently when using categorical data, even worse for mixed data.
CARTMVI [34] (✓)	Seeks predictors and cut points in the predictors that are used to split the sample. The cut points divide the sample into more homogeneous subsamples.	Can handle numerical data that are highly skewed or multi-modal and categorical predictors with either ordinal or nonordinal structure.
C4.5MVI [35] (✓)	C4.5 performs pruning by replacing a subtree in the decision tree with a single decision node that contains all the subtree decisions.	Unstable, meaning that a slight change in the data can lead to a significant shift in the structure of the optimal decision tree and requires more memory.
ELMMVI [36] (✓)	Consists of three layers: input, hidden, and output layers, based on empirical risk minimization theory; avoids multiple iterations and local minimization.	Faster than most existing neural network algorithms, as it is a single hidden layer feedforward network and yields promising performance.
GAMVI [37] (✗)	It is introduced to generate optimal sets of missing values, and information gain is used as the fitness function to measure individual imputation's	Requires fewer data about the problem, but building an objective function and getting the suitable operators can be difficult and computationally expensive.

	performance.	
GKNNMVI [38] (✓)	Based on the grey coefficient, the valid attribute values derived from these nearest neighbors are used to predict the missing values.	It gives a normalized measure function, and mutual information-based feature relevance is embedded during estimating missing values
IKNNMVI [39] (✓)	Imputes missing values by K-nearest neighbors via calculating the gray distance between the missing datum rather than traditional Euclidean distance methods.	Searches for the nearest neighbor instance with the same class label between the instance and the missing instance, reducing the time complexity with reduced errors.
SKNNMVI [47] (✓)	Fills by taking the mean of the NN genes in the complete set where Euclidean Distance is used as the distance metric to determine the NN genes.	Has improved accuracy and computational complexity over the conventional KNN-based and other maximum likelihood estimation-based methods.
WKNNMVI [40] (✓)	Weighted KNN is employed, where weight is the correlation between a missing dimension and available data values from other fields.	Performance depends on data quality and is sensitive to the scale of data, which is also computationally expensive.
KMVI [41] (✗)	Imputes missing values by drawing from the posterior predictive distribution of the missing data given the observed data.	Convergence is very slow and not applicable when confidence intervals of estimated parameters are the primary goals.
MLPMVI [42] (✓)	Construct an MLP model, feed the data without having any missing value and finally predict for the missing value instances.	Many MLP models have to be constructed when missing items appear in several attributes in a high-dimensional problem
RFMVI [43] (✓)	A tree is grown from randomly selecting training samples, as much as possible without pruning, where the top trees' features are more important than end nodes.	Found to be biased while dealing with categorical variables. Many trees can make the algorithm too slow and ineffective for real-time predictions.
RSTMVI [44] (✓)	The rough set rule induction method enables obtaining association rules of missing data patterns based on approximations, dependencies, and decision rules.	Does not need any preliminary or additional information, such as a probability distribution of the given data samples.
SVRMVI [45] (✓)	Estimates parameters directly, minimizing the distance (max. distance and Hausdorff distance) between the actual and prediction, using a ϵ-SVR model.	Suited for only interval data and reliant on the outliers and classes overlapping. It also underperforms for more features with fewer training samples.

Despite introducing a novel MVI method or being chosen for the specific field of experiments, some of the MVI methods listed in Table 1 are reported as the basis for pertinent experiments, such as medical and financial decision-making systems, pattern classification, questionnaires, and industrial operation management. It is necessary to emphasis thatpaper's main objective is not to describe the concise theories underlying various MVI approaches. As a result, we provide the citations for the respective MVI methods in Table 1 so that readers can study and learn deep theoretical facts,.

The majority of early polls focused on outlining the essential concepts of MVI systems, which were the most popular and MVI's repercussions on ML-based decision-making systems are not thoroughly evaluated and described. This article seeks to present a thorough analysis of the MVI methodologies along with other crucial pertinent studies. However, Table 2 lists all of the statistical and machine learning (ML) based MVI systems that were commonly used from 2010 to August 2021 in various published studies. This discovery offers a preliminary estimation to researchers at all levels, from novice to expert, to help them select an appropriate MVI algorithm for their decision-making pipeline. It is noticeable from Table 2 that a few algorithms, including EMMVI, HDMVI, LLSMVI, LRMVI, MMVI, MICEMVI, BPCAMVI, SVDMVI, ANNMVI, KMCMVI, FCMMVI, CARTMVI, KNNMVI, and RFMVI, are most regularly exercised in the last decade (2010–August 2021), compared to all other algorithms in Fig. 4. Table 2 tells that statistical MVI, such as EMMVI and MMVI, and ML-based MVI, e.g., KNNMVI and RF, are massively employed imputation methods in the last decade.

Table 2 The statistical and ML-based techniques, applied in the literature. The publications have been selected

Methods	Related publications
Statistic-based MVI methods	
EMMVI	Aussem and de Morais [48], Ghannad-Rezaie et al. [49], Hron et al. [10], Jerez et al. [42],
GMMMVI	García-Laencina et al. [8], Kang [28],
HDMVI	Jerez et al. [42], Ghorbani and Desmarais [48]
MCMCMVI	Ding and Ross [46], Ghannad-Rezaieet al. [49].
ILLSMVI	Hron et al. [10], Pati and Das [47].
LTSMVI	Hron et al. [10].
LLSMVI	Luengo et al. [43],
SBEMVI	Jahan et al. [12].
LRMVI	Jerez et al. [42], Silva-Ramírez et al. [7], Jahan et al. [12]
TMVI	Ghannad-Rezaieet al. [49],
IDMVI	Liu et al. [14].
MMVI	Hron et al. [10], Jerez et al. [42], Silva-Ramírez et al. [7], Wei et al. [25], Xu et al. [19],
MCMVI	Thulare et al. [28].
MICEMVI	Jerez et al. [42], Valdiviezo and Van Aelst [18],
MISRTMVI	Valdiviezo and Van Aelst [18].
MIADMVI	Xu et al. [19].
BPCAMVI	Pati and Das [47]
PCAMVI	Malan et al. [5]
CCMVI	Tsai et al. [23], Brahmaet. al. [52]
GRLSRMVI	Chen et al. [35].
QRILCMVI	Wei et al. [25].
SVDMVI	Wei et al. [25]
ML-based MVI methods	
ANNMVI	Choudhury and Pal [27]
ARMVI	Vougas et al. [14]
KMCMVI	Migdady and Al-Talib [31]
SOMMVI	Singh et al. [33], Jerez et al. [42]
CARTMVI	Purwar and Singh [1], Ghannad-Rezaie et al. [49]
GAMVI	Lin and Tsai [2]
KNNMVI	Valdiviezo and Van Aelst [18], Hron et al. [10], Jerez et al. [42].
GKNNMVI	Zhang [39].
SKNNMVI	Pati and Das [47]
MLPMVI	Choudhury and Pal [27], Silva-Ramírez et al. [7], Jerez et al. [42]
RFMVI	Kokla et al. [43], Wei et al. [25], Purwar and Singh [1], Valdiviezo and Van Aelst [18],

In the past ten years, statistical MVI methods, such EMMVI and MMVI, as well as ML-based MVI methods, like KNNMVI and RF, have seen widespread use. The current parameter is determined by the E-step of the EMMVI algorithm from all available data, and it is modernised by the M-step by maximising the likelihood function. The missing values are determined from the revised probability function after the process has continued up to the stopping condition. The mean (average), median (middle), or mode (most common) values of an attribute from the entire dataset are used to estimate the missing values in the MMVI technique for the MVI. In the KNNMVI policy, the measured values from the k nearest observed values are used to fill in the missing values using a distance function, often the Euclidean distance.

Missing data is used as the testing step, where both the complete and missing attributes represent the input features and provide individual class labels. Using the bootstrapping process, different decision trees are put together in the RFMVI technique. The final forecasts are given by the averaged values or majority votes of each tree's prognostication. The inner and leaf nodes of the KNNMVI imputation algorithm, which RFMVI uses, respectively define the inputs attributes and the outputted class labels. The metrics for MVI method evaluations and comparisons are discussed and studied in the section that follows.

3. Imputation Evaluation

The next step is to evaluate the outcomes of the imputation once the missing values have been imputed using one of the MVI methods, as explained in the earlier section. This can be done in two ways: directly and indirectly by classification accuracy. Again, the attributes primarily have two types of values: discrete or categorical and continuous. By evaluating the Percentage of Correct Predictions (PCP) [2, 18] and Cross-Entropy (CE) (see their mathematical definitions in Table 3), the former categorical value imputations are typically directly evaluated. The Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) [2, Root Mean Square Error (RMSE) [2, RMSE Improvement (RMSEI), and Similarity metrics are computed to directly analyze the later continuous value imputations. Angle (SA), Similarity Length (SL), Similarity Fraction of the Same Neighbors (SFSN), Pearson Correlation Coefficient (PCC) , $R2$ [7], and Mean Confidence Interval Length (MCIL) (see their mathematical definitions in Table 3). Again, many researchers have estimated the modified RMSE (mRMSE) value √ from both the categorical and continuous attributes as$\mathrm{mRMSE} = \sqrt{RMSE_0^2 + RMSE_1^2}$, where $RMSE_0$ and $RMSE_1$ are RMSEs of categorical and continuous variables, respectively and are defined in Eq. (2).

$$\mathrm{RMSE}_0 = \sqrt{\frac{1}{|M_0|}} \textstyle\sum_{(i,j)\in M_0} 1_{\{\hat{x}_{ij} \neq x_{ij}\}},$$

$$\mathrm{RMSE}_1 = \sqrt{\frac{1}{|M_1|}} \textstyle\sum_{(i,j)\in M_0} 1_{(\hat{x}_{ij} - x_{ij})^2}, \qquad (2)$$

where the evaluation's "ground truth" in X is represented by the known values "xij" (i, j) "M0," and "M1." The RMSE metrics were also adjusted by the authors in a number of studies, including Normalised RMSE (NRMSE) and Coefficient of variation RMSE (CVRMSE). While the latter CVRMSE is obtained by dividing the computed RMSE by the mean of the actual values, the former NRMSE is obtained by normalising the calculated RMSE value. The RMSE number can also be used to calculate the Mean Square Error (MSE) [7] and the Normalised MSE (NMSE) in various studies. The normalised form of the MSE is called NMSE, and the MSE is calculated as MSE = RMSE2. The time needed by the algorithm to impute the missing values, whether categorical or continuous, is measured using the Total Computation Time (TCT) metric, which is also used to assess the MVI approaches. The size of the dataset and the incidence of missing data determine the necessary TCT.

The alternative indirect evaluation technique looks at the classification metrics of a few selected classifiers that were trained using the outputted whole dataset with imputed values (further information in Section 3.2). In contrast to the direct evaluation method, some specific classification models are trained using the imputed dataset (see detailed models in Section 3.1). The classifier that produces better results for the same metric implies that the dataset it uses as input has greater imputation quality, which leads to better imputation techniques being produced. But Table 4 shows how those 191 papers were categorised based on their direct and indirect MVI evaluation, demonstrating that direct approaches are more active in comparing indirect assessment.

Table 3 clearly shows that direct assessment procedures have been heavily used for missingness imputation throughout the past ten years (2010–August 2021). The meticulous examination of each of the 191 articles that were chosen reveals that there are much fewer works that use both assessment policies, which correspond to the direct and indirect evaluation procedures.

4. ML models and their evaluation

This section briefly introduces various ML models in Section 2.1 and the evaluation metrics for the ML models in Section 2.2. The survey and the employment of those two items in selected articles from 2010 to 2021 are also presented in this section.

4.1 ML model

In recent years, data analysis and computers have seen an increase in the use of artificial intelligence (AI), particularly machine learning (ML), which typically enables programs to perform intelligently. ML is typically referred to as the most well-liked new technology in the fourth industrial revolution and gives systems the ability to learn from experience and improve without having precisely programmed automatically[50,52,53,54,55].

It has also been used in the digital sphere, including the internet of things, cybersecurity, mobile data, business, social media, health data, etc.. The four main subtypes of machine learning algorithms are reinforcement learning, semi-supervised learning, unsupervised learning, and supervised learning.The ML model is typically responsible for supervised learning, also known as the task-driven technique, to learn a function that translates an input to an output based on sample input-output pairs.

To comprehend a function, it makes use of labeled training data and a collection of training examples. Classification and regression are the two most typical supervised tasks. The first task categorises the data, while the second regression task fits the data. Without the need for human intervention, unsupervised learning analyses unlabeled datasets and is frequently used to identify generative features, discern important patterns and compositions, group results, and generate experimental ideas. Clustering, density estimation, feature learning, dimensionality reduction, association rule discovery, and anomaly detection are the most common unsupervised learning tasks. A hybridization of the previously mentioned supervised and unsupervised procedures is what is known as semi-supervised learning. As it utilizes both labeled and unlabeled data, as stated above, as a result, the choice is between learning under supervision and learning independently. Semi-supervised training is advantageous in real-world circumstances where labeled data may be scarce in sparse settings while unlabeled data are abundant. A semi-supervised learning model's ultimate objective is to provide a prognostication result that is more favourable than one obtained by using only the labelled data from the model. Machine translation, fraud detection, data labelling, and text classification are some job-related applications of semi-supervised learning. An ML approach called reinforcement learning enables software tools and machines to automatically analyse the ideal performance in a particular connection for enhancing productivity, i.e., an environment-driven process.

This kind of training is focused on rewards or penalties, and its ultimate goal is to use the wisdom gained from environmental activists to enhance the rewards or reduce the uncertainty. It is a powerful tool for developing AI models that promote increased automation or enhance the efficiency of well-established processes like robotics, autonomous driving, manufacturing, and supply chain logistics. To determine the fundamental or simple problems, though, is not the best use of it. These debates come to the conclusion that the five categories of ML models—Classification, Regression, Clustering, Association rule learning, and Reinforcement learning—can be grouped together[51]. The research papers that were chosen for this review are shown in Fig. 5, where we provide all five of the aforementioned groupings. The first category, classification, is a learning method to a predictive modeling problem, where a class label is predicted for a given example. The second category, regression analysis, uses five alternative models to forecast a continuous (y) result variable based on the value of one or more (x) predictor variables

After carefully examining those selected articles, Fig. 6 presents the top 15 operational ML models. This graph clearly shows that KNN, which scored three times higher than the second-highest RF model, is the best method for classifying the 191 articles. Fig. 6 further shows that SVM, BPCA, and DT, which are the third, fourth, and fifth-most used ML models, respectively, are used in approximately the same number of papers. The KNN model is widely used in the literature for a number of plausible reasons, including its quick calculation time, straightforward method to interpret, versatile utility for regression and classification, and ability to change with fresh data.

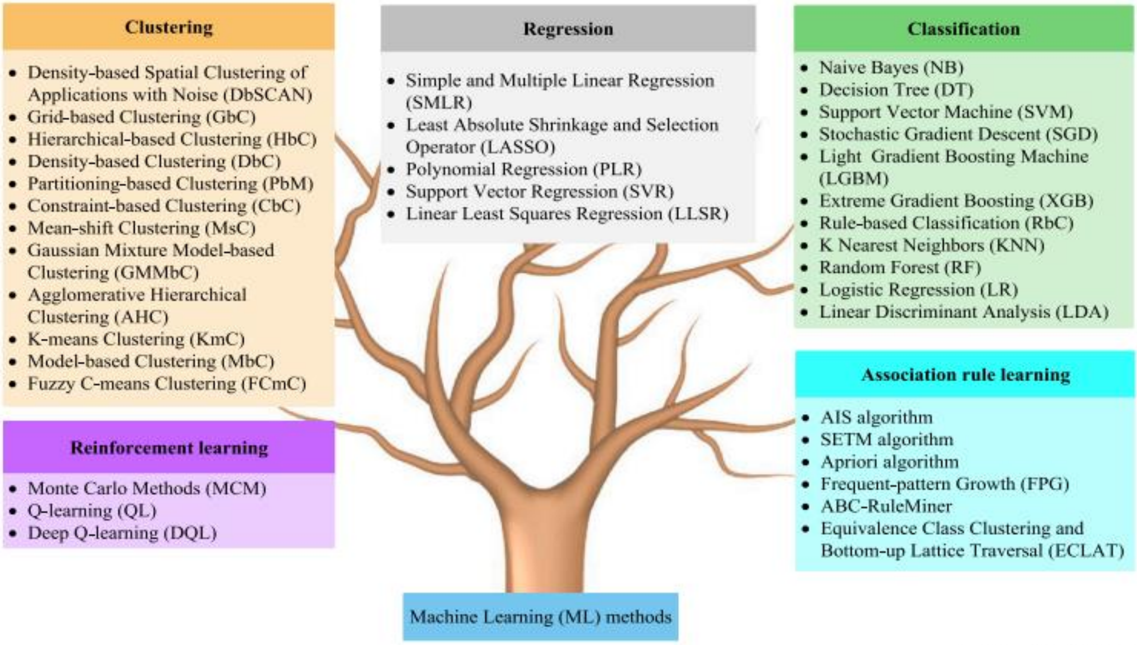

Clustering
• Density-based Spatial Clustering of Applications with Noise (DbSCAN)
• Grid-based Clustering (GbC)
• Hierarchical-based Clustering (HbC)
• Density-based Clustering (DbC)
• Partitioning-based Clustering (PbM)
• Constraint-based Clustering (CbC)
• Mean-shift Clustering (MsC)
• Gaussian Mixture Model-based Clustering (GMMbC)
• Agglomerative Hierarchical Clustering (AHC)
• K-means Clustering (KmC)
• Model-based Clustering (MbC)
• Fuzzy C-means Clustering (FCmC)
Regression
• Simple and Multiple Linear Regression (SMLR)
• Least Absolute Shrinkage and Selection Operator (LASSO)
• Polynomial Regression (PLR)
• Support Vector Regression (SVR)
• Linear Least Squares Regression (LLSR)
Classification
• Naive Bayes (NB)
• Decision Tree (DT)
• Support Vector Machine (SVM)
• Stochastic Gradient Descent (SGD)
• Light Gradient Boosting Machine (LGBM)
• Extreme Gradient Boosting (XGB)
• Rule-based Classification (RbC)
• K Nearest Neighbors (KNN)
• Random Forest (RF)
• Logistic Regression (LR)
• Linear Discriminant Analysis (LDA)
Reinforcement learning
• Monte Carlo Methods (MCM)
• Q-learning (QL)
• Deep Q-learning (DQL)
Association rule learning
• AIS algorithm
• SETM algorithm
• Apriori algorithm
• Frequent-pattern Growth (FPG)
• ABC-RuleMiner
• Equivalence Class Clustering and Bottom-up Lattice Traversal (ECLAT)
Machine Learning (ML) methods

Fig. 5. Tree diagram of different ML models, where we group them into five categories depending on their similarity.

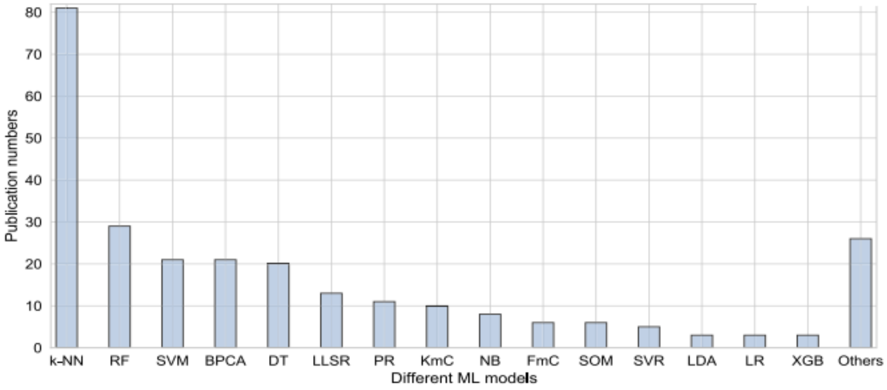
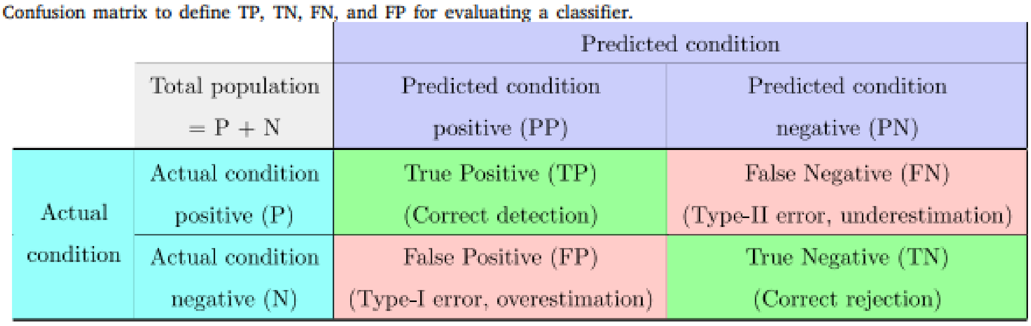


Fig. 6. Illustration of the number of articles on fifteen different ML models in the selected 191 articles, where the deep learning and multi-layer perceptron are included in the other category.

Table 5
Confusion matrix to define TP, TN, FN, and FP for evaluating a classifier.



		Predicted condition	
	Total population = P + N	Predicted condition positive (PP)	Predicted condition negative (PN)
Actual condition	Actual condition positive (P)	True Positive (TP) (Correct detection)	False Negative (FN) (Type-II error, underestimation)
	Actual condition negative (N)	False Positive (FP) (Type-I error, overestimation)	True Negative (TN) (Correct rejection)

4.2. ML classifier evaluation

As shown in Table 5, a confusion matrix, which contains information about actual and anticipated classes and establishes the essential associations to understand accuracy computations for a certain classifier, is typically used to track the performance of classifiers. Each column in that matrix (see Table 5) denotes potential outcomes, whereas each row gives concrete examples. When a condition is present, the proper anticipations, such as True Positive (TP) and True Negative (TN), respectively, show that the disease has been correctly recognised. When an infection is absent, however, the sickness has been effectively rejected. False negatives (FN), also known as false-negative errors, are test results that incorrectly claim a condition does not exist. A False Positive (FP) or false-positive error, on the other hand, is a determination that a particular condition exists when it actually does not. The metrics True-positive Rate (or Sensitivity, or Recall), True-negative Rate (or Specificity, or Selectivity), True-positive Rate (or False alarm, or Fall-out), False-negative Rate (or Miss rate), False-negative Rate (or Miss rate), Positive Predictive Value (or Precision), Negative Predictive Value (NPV), False Discovery Rate (FDR), False Omission Rate (FOR), Accuracy, and F1 score, Receiver Operating Characteristic (ROC) curve, Area under the ROC curve (AUC), Precision–recall (PR) curve, Area under the ROC curve (AUC), Precision–recall (PR) curve, and Area under the PR curve (AP) (see their mathematical details in Table 6).

The multi-class evaluation metrics are expansions of the binary metrics (see Table 6), which average the metrics over all classes in a variety of ways, including micro, macro, etc. Class imbalance is taken into account by the former averaging approach (micro), which totals the instances each class was properly and wrongly predicted to calculate the metrics. The latter strategy (macro), in contrast, measures each class's metrics independently and determines their not weighted mean without taking imbalance into account. K-fold cross-validation is occasionally used by researchers to confirm the robustness of the trained model. In this situation, the metrics that were received from each fold are expressed as an average value with a standard deviation, or Eq. (3). Increased standard deviation values indicate increased inter-fold variance, which in turn denotes inadequate robustness, and vice versa.

$$M = \frac{1}{K} \times \sum_{i=1}^{K} M_i \pm \sqrt{\frac{\sum_{i=1}^{K}(M_i-\mu)^2}{K}} \quad (3)$$

where $Mi \in$ R, $i \forall K$, denotes an estimated metric with a mean value of μ and K is the fold numbers. Accuracy, sensitivity, and specificity are the most often used metrics for evaluation, as revealed by the literature's subsequent one by one examination. These measures are used in literature by 39.13%, 11.96%, and 14.13 %, respectively. Although accuracy is frequently used as a statistic, class imbalance is not taken into account. The ML models have a tendency to learn a class more precisely if there are a lot of samples in that class, which raises the false-positive or false-negative rates. If there is no penalization of class imbalance, which is taken into consideration in some measures, such as ROC and AUC estimation, the conventional metrics provide very high values. yet, just 8.70% of attempts.

5. Discussion

This part provides an example-based analysis of several examined methods from the aforementioned sections, highlighting certain flaws related to technical issues with the experimental design that are thought to be future research goals for the MVI approaches (see part 4.1). Additionally, we investigate how the MVI approach might be used to ML classifiers in order to improve the results (see Section 4.2). Additionally, we present illuminating recent research trends on the approaches under discussion, compiling data from the previous 10 years from the 191 publications gathered, as well as some suggestions for future study directions (see Section 5.3).

5.1 Analysis of MVI methods and their evaluation

Figure 7 bestows the top twelve highly applied statistical and ML based methods for the MVI, commonly employed in the literature from 2010 to 2021. This graph shows that the top five statistical MVI techniques—EMMVI, MMVI, LLSMVI, BPCAMVI, and LRMVI—are consistently chosen and used in 34, 34, 12, 11, and 11, respectively, articles. It is interesting that the statistical EMMVI and MMVI are used more frequently than the other two MVI algorithms (LLSMVI), with their use being almost three times higher. These two approaches, which were also the results of a survey conducted from 2006 to 2017 by Lin and Tsai [2], should be regarded as the typical baseline statistical MVI procedures, in accordance with our study. These could be the justifications for using those techniques: a small amount of time is required for the prediction of missing values, they are simple to implement and memory-efficient, less affected by the outlier imputation as they look for values within the ranges of the attributes, requires no prior knowledge of the data, and imputed values maintain a nearly unbiased mean for the attributes. Although the EMMVI or MMVI imputation maintains the attribute's mean, they do not take into account the co-relation between the attributes, which may negatively impact the missingness imputation in some applications, particularly if the mutual relations of the attributes are important. Once more, the MMVI is not an appropriate MVI technique for categorical attribute values since it may produce fractional values that must be trimmed with the missing data. The third and fifth most popular models, LLSMVI and LRMVI, on the other hand, estimate relationships between characteristics before determining the missing values using the regression coefficients. These two types of MVI, LLSMVI and LRMVI, are used in practice to predict numerical and categorical attribute values, as reported in many articles of the chosen 191 papers.

In addition, the top five ML-based MVI approaches KNNMVI, RFMVI, KMCMVI, FCMMVI, and CARTMVI have all been utilised in 24, 12, 7, 7, and 7 articles, respectively. As it is around two times more effective than the second MVI algorithm (RFMVI), which may be regarded as the usual representative baseline ML-based MVI approach, the KNNMVI is the most widely used ML-based MVI procedure. KNNMVI, RFMVI, and CARTMVI, three of the top five MVI methods, are supervised approaches, whereas KMCMVI and FCMMVI are unsupervised systems. In the previous supervised systems, the entire dataset served as the training set while the missing dataset served as the testing set. The latter unsupervised clustering techniques, in contrast, combine a collection of related items into identical assemblages. Remarkably, the mean of the objects in a cluster form the individual cluster center.

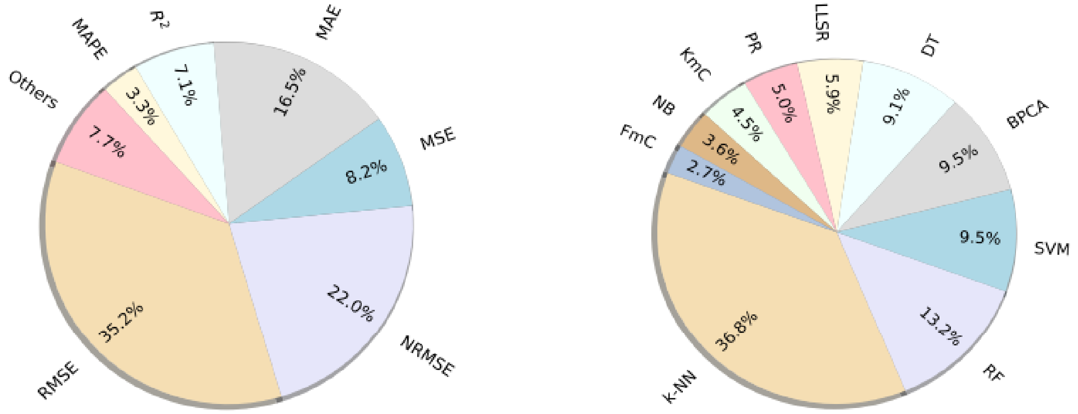
The nearest centroid's values are used to fill in any missing values after determining the distance between the incomplete data and the recognised cluster centroids. If the starting guesstimation is too far off from the actual answers, such cluster algorithms require huge iterations in order to reach the stopping criterion. On the other hand, supervised KNNMVI has a quick computation time, a straightforward, understandable methodology, and it can adapt to new data, which may be the legal reasons for the KNNMVI method's widespread adoption. Finally, it is interesting that statistical MVI techniques have been used more frequently in the recent ten years compared to ML-based MVI systems based on our survey. Once more, Fig. 7 shows that the top 12 statistical MVI systems are all used more frequently than their comparable ML-based MVI counterparts. The statistical MVI techniques can be favoured because they don't call for training on powerful machines and quick imputation. Evaluation is especially important for confirming the performance of the MVIs and providing the final results. The quantity of publications for direct and indirect evaluation against each year is provided by the summary of the MVI techniques' evaluation metrics in Table 4 in Fig. 8. This graph shows that the direct evaluation methods are heavily used through some metrics for all years except 2012 (see Table 3). Additionally, it should be noted from Fig. 8 that the MVI process assessment heavily relies on direct evaluation methods as of 2017. As shown in Fig. 8, the percentage of pertinent tasks that use an indirect assessment policy is significantly lower than the percentage that uses the direct judgement approach. The article numbers also take into both evaluation procedures at the same time are indeed negligible, as also reviewed by Lin and Tsai [2]. [The selected paper's additional investigation reveals the most frequently used direct measurements and classification models for the MVI policy assessment, which are shown in the pie plot in Fig. 9. The most often used direct measures, according to the left figure in Fig. 9(a), are RMSE, NRMSE, MSE, MAE, R2, and MAPE. Only 7.7% of articles published in the past ten years used the criteria in Table 3 that weren't already stated. Additionally, it should be noticed that the RMSE and NRMSE measures used in the 57.2% of papers are similar. The RMSE measure is therefore a far superior evaluation criterion for measuring the last ten years' trends. The RMSE measure is a much better evaluation criterion for assessing the correctness of the data missingness imputation than the trends of the previous decade. Again, Fig. 9(b) displays the proportion of the top ten most popular ML models for indirect evaluation, including KNN, RF, SVM, BPCA, DT, LLSR, PR, KmC, NB, and FCmC. From Fig. 9(b), it is striking to see that KNN and RF models have been the most often used ML models for indirect assessment during the past ten years, accounting for about 50% of all articles from 2010 to 2021.

5.2. Effect of MVIs on ML models

We select two distinct datasets for two distinct applications with missing values in Sections 2.4.2.1 and 2.4.2.2, and we analyse the accompanying articles to ascertain the impact of the MVI on the diagnostic or decision-making pipelines of these applications. In addition, eleven different datasets from various fields are provided in Section 2.4.2.3, showing the essence of MVI and gathered from several previously published works.

5.2.1PIMA Indians Diabetes (PID) dataset

768 female diabetic patients from the Pima Indian community in the vicinity of Phoenix, Arizona, are included in the PIMA Indians Diabetes (PID) dataset [328]. The 268 diabetic patients (positive) and 500 non-diabetic patients (negative) in this PID dataset each have eight features. The statistical breakdown and description of this PID dataset are shown in Table 7. F3, F4, and F5 in Table 7 each have a distinct number of missing values, such as 35, 227, and 374, out of the eight total attributes. In the paragraph that follows, we look into five distinct articles to see how the use of MVIs has improved performance.



(a) Metrics for direct MVI evaluation

(b) ML models for indirect MVI assessment

Fig. 9. Pie plot of different metrics and ML models for evaluating MVI methods, dispensing their utilization percentage in the selected articles.

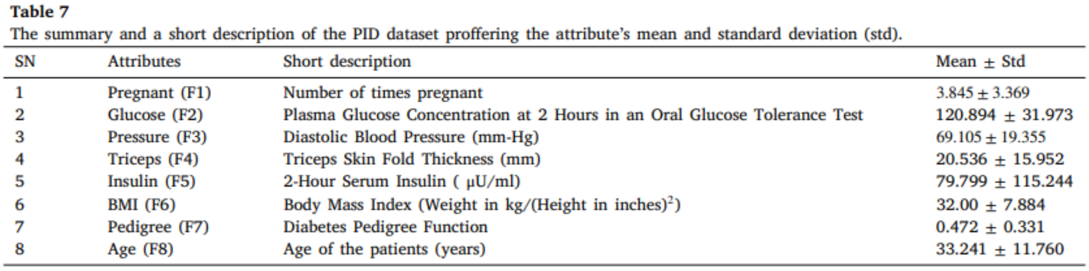


Table 7
The summary and a short description of the PID dataset proffering the attribute's mean and standard deviation (std).

SN	Attributes	Short description	Mean ± Std
1	Pregnant (F1)	Number of times pregnant	3.845 ± 3.369
2	Glucose (F2)	Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test	120.894 ± 31.973
3	Pressure (F3)	Diastolic Blood Pressure (mm-Hg)	69.105 ± 19.355
4	Triceps (F4)	Triceps Skin Fold Thickness (mm)	20.536 ± 15.952
5	Insulin (F5)	2-Hour Serum Insulin (μU/ml)	79.799 ± 115.244
6	BMI (F6)	Body Mass Index (Weight in kg/(Height in inches)^2)	32.00 ± 7.884
7	Pedigree (F7)	Diabetes Pedigree Function	0.472 ± 0.331
8	Age (F8)	Age of the patients (years)	33.241 ± 11.760

In their suggested pipeline, the authors trained various ML models with or without MVI, including KNN, DT, RF, AdaBoost (AdB), NB, and XGB. Their experimental findings showed that, when the XGB classifier was used, the addition of MVI increased the AUC values by a factor of 10.9%. In their suggested DMP_MI pipeline, using 5-fold cross-validation to compare their imputation method to an RF classifier alone, the authors were able to demonstrate that it increased the AUC value by 5.8%. Christobel and SivaPrakasamintegrated a supervised KNN-based imputation (i.e., KNNMVI) with the KNN classifier, providing a 1.5% increase in accuracy as a result of the imputation of missing values. The RF classifier with group median imputation (i.e., MMVI) and other methods give an enhanced accuracy of 1.0%, according to the authors. In Fig. 10, which illustrates the effectiveness of MVI method application, the summary of those publications for the performance enhancements of the ML model as a result of the MVI method employment is shown.

5.2. Heart Disease (HD) dataset

The Heart Disease (HD) dataset, which has 303 observations in two classes (sick with 164 samples and normal with 139 samples), is accessible in the UCI Machine Learning Repository]. The thirteen features of this HD dataset are summarised statistically and are described in detail in Table 8. Table 8 contains a total of thirteen qualities, and of those, F7, F8, F10, and F11 each have varying numbers of missing data (21, and 2, respectively). In the paragraph that follows, we examine five different articles on the HD dataset to evaluate the performance gains brought on by the use of MVIs.

By combining two separate MVI techniques—ML-based KNNMVI and statistical HDMVI—with a fuzzy SVM classifier and additional preprocessing, Nilashi et al. were able to increase accuracy by 2.5%.

With a 1.1% advantage over HDMVI, the KNNMVI also exceeded it in terms of accuracy for their suggested pipeline. Again, Khennou et al. used the ML-based KNNMVI approach with an SVM model, producing results that were 11.6% more accurate than they would have been without imputation. For this dataset, many authors used ANNMVI with KNN classifier, which resulted in accuracy gains of 5.6% as a result of missing value imputation.Researcher introduced a hybrid classifier with weighted voting, and they statistical MMVI to impute the missing data.

They conducted experiments to demonstrate how the addition of MMVI to their classifier increased accuracy by a margin of 6.3%. which demonstrates the efficiency of MVI process employment are primarily employed, which is the similar outcomes of Fig. 6. It is also remarked that the statistical MVI methods are widely applied for the missingness imputation, which essentially enhances the prediction results.

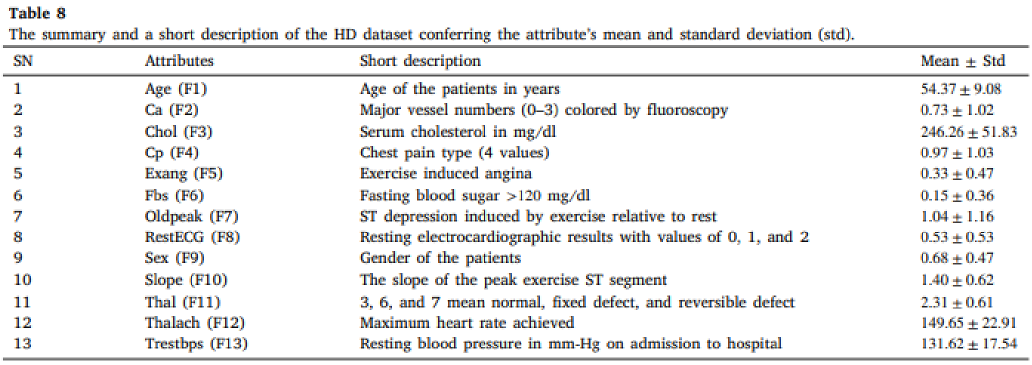


Table 8
The summary and a short description of the HD dataset conferring the attribute's mean and standard deviation (std).

SN	Attributes	Short description	Mean ± Std
1	Age (F1)	Age of the patients in years	54.37 ± 9.08
2	Ca (F2)	Major vessel numbers (0–3) colored by fluoroscopy	0.73 ± 1.02
3	Chol (F3)	Serum cholesterol in mg/dl	246.26 ± 51.83
4	Cp (F4)	Chest pain type (4 values)	0.97 ± 1.03
5	Exang (F5)	Exercise induced angina	0.33 ± 0.47
6	Fbs (F6)	Fasting blood sugar >120 mg/dl	0.15 ± 0.36
7	Oldpeak (F7)	ST depression induced by exercise relative to rest	1.04 ± 1.16
8	RestECG (F8)	Resting electrocardiographic results with values of 0, 1, and 2	0.53 ± 0.53
9	Sex (F9)	Gender of the patients	0.68 ± 0.47
10	Slope (F10)	The slope of the peak exercise ST segment	1.40 ± 0.62
11	Thal (F11)	3, 6, and 7 mean normal, fixed defect, and reversible defect	2.31 ± 0.61
12	Thalach (F12)	Maximum heart rate achieved	149.65 ± 22.91
13	Trestbps (F13)	Resting blood pressure in mm-Hg on admission to hospital	131.62 ± 17.54

5.3. Observations and recommendations

This section reveals our critical observations, obtained from point-to-point scrutinization of the selected 191 articles thoroughly. Consequently, we point out several recommendations for the imputation of missing value(s), which will provide future direction for novice to seasoned researchers building their framework for generating decisions from the incomplete datasets. The following are noteworthy findings and suggestions:

- The importance of the missed value(s) will determine if the MVI process is required. There are a number of cutting-edge classification or regression techniques that can manage data missingness on their own internally and do not require the MVI method to handle the incomplete datasets. Beginning with the incomplete and complete datasets (imputed by the indicated representative baseline MVI method(s) in this article), the researcher can derive approximative classification results from those models using the incomplete and complete datasets. Such an experiment might highlight the significance of the missed value(s) and, as a result, demonstrate the need for missingness imputation or not.
- It should be noted from the review and analysis of this article that, regrettably, there is no clear answer to the query, "Which method is best for the missingness imputation?" Since the MVI approaches depend on different constituents of interest, researchers must look for the best approach. The investigation of the 191 articles that were chosen as well as the additional 21 articles (in Section 4.2), where the authors used various MVI techniques to improve the performance of ML models, reveals that the adoption of MVI methods depends on a variety of factors or alternative techniques, including attribute selection, outlier detection, attribute normalisation, the classifier(s), the field of application(s), computational resource availability, time to receive imputation, and object factor consideration (either mean conserving or inter-attributes' correlation preserving).
- Since attribute and/or sample selection tries to eliminate unrepresentative attributes and/or samples from the input feature vector, its use before or after the MVI process may have an impact on the imputation results. If one or both of these tasks were completed prior to the MVI process, the learning phase's whole dataset would be cleaner, and the imputation returns would likely be more reliable. Alternately, using attribute and/or sample determination over an imputed dataset after MVI may result in a classifier that is more advantageous than one based just on the imputed dataset.
- Although the imputation process rarely depends on the imputation approach, the normalisation or standardisation of the data can have a significant impact. For instance, it is advised to standardise the data before impute it when using distance-based algorithms (such as KNNMVI); the lower values of the attributes converge more quickly and need less computation when used in real-time. There is debate regarding whether to standardise or impute first. If standardisation is carried out initially, the imputation procedure might have an impact on the choice of centre and scale. On the other hand, using imputation first helps lessen the skewness of the estimated mean and scale brought on by the pattern's missingness.

- Outliers can make the imputed values occasionally unreliable, which has a negative impact on the values entered for missing data. The management of outliers must therefore be done before missingness imputation. Surprisingly, the outliers have a significant negative impact on regression-type MVI approaches, necessitating outlier rejection. On the other hand, because they impute the missing values from the most common values, which might not be outliers, median value-based MVI approaches can reduce the effects of outliers.
- The three most popular methods for evaluating the MVI method(s) are the direct assessment system, the defined metric(s) of the classifiers in indirect evaluation, and reflection of TCT. All three of these evaluation methodologies should be utilised in order to fully explain the performance of the MVI technique(s) and offer recommendations for creating more dependable imputation methods. Unfortunately, it has only sometimes been used for assessment in related studies over the past ten years. This is one of the problems with the papers' existing methods, and the researcher should take them all into account in further experiments.

6. Conclusion

In data mining, big data analysis, and ML-based decision-making pipelines, the MVI method for incomplete datasets is a critical concern since the final mining or analysis result could be negatively affected if the missing datasets are not properly imputed. This article reviews and investigates 191 relevant publications that were published between 2010 and 2021. This article's assessment and analysis concentrate on the problems encountered during the MVI process, including the MVI techniques used and the evaluation schemes considered. The analysis of the publications over the previous ten years reveals a number of issues with the literature, including the best methods for MVI and its evaluation, the elements that can negatively impact missingness imputation, and the influence of MVI methods on the decision-making process.The investigated conclusions from the selected 191 MVI’s articles reveals that EMMVI, HDMVI, LLSMVI, LRMVI, MMVI, MICEMVI, BPCAMVI, SVDMVI, ANNMVI, KMCMVI, FCMMVI, CARTMVI, KNNMVI, and RFMVI, are the most regularly practiced MVI strategies in the last decade.The findings also show that statistical MVIs like EMMVI, MMVI, LLSMVI, BPCAMVI, and LRMVI are more widely used approaches because they don't call for special training on complex machines and quick imputation. Additionally, the top five metrics for the direct MVI evaluation are RMSE, NRMSE, MSE, MAE, and R2. In contrast, the top five ML models used for indirect MVI evaluation are KNN, RF, SVM, BPCA, and DT. The results of this paper should be useful for recovering those issues, choosing an appropriate MVI approach, and choosing its evaluation metric for the associated research community. The suggestions could also be a great directive for future researchers to create an efficient decision-making system utilising ML model(s) with sparse datasets for several real-world applications.

References

[1] PurwarA,SinghSK.Hybridprediction model with missing value imputationformedicaldata.ExpertSystAppl2015;42:5621–31.

[2] LinWC,TsaiCF.Missingvalueimputation:areviewandanalysisoftheliterature(2006–2017).ArtifIntellRev2020;53:1487–509.

[3] KhalidM,SinghGN.Someimputationmethods to deal with the issue ofmissing data problems due to random non-response in two-occasion successivesampling.CommStatistSimulationComput2020;1–21.

[4] IslamMR,MoniMA,IslamMM,Rashed-Al-MahfuzM,IslamMS,HasanMK,Hossain MS, Ahmad M, Uddin S, Azad A, et al. Emotion recognition from EEGsignal focusing on deep learning and shallow learning techniques. IEEE Access2021;9:94601–24.

[5] Rahman MG, Islam MZ. iDMI: A novel technique for missing value imputationusingadecisiontreeandexpectation-maximizationalgorithm. In: 16th Int’lconf.computerandinformationtechnology.IEEE;2014,p.496–501.

[6] YanX,XiongW,Hu L, Wang F, Zhao K. Missing valueimputation

basedongaussianmixturemodelfortheinternetofthings.MathProblEng2015;2015:1–8.

[7] Silva-Ramírez E-L, Pino-MejíasR, López-CoelloM, Cubiles-de-la Vega M-D.Missingvalueimputationonmissingcompletelyatrandomdatausingmultilayerperceptrons.NeuralNetw2011;24:121–9.

[8] SuhaimiN,GhazaliNA,NasirMY,MokhtarMIZ,RamliNA.MarkovchainMonteCarlomethodforhandlingmissingdatainairqualitydatasets.MalaysJAnalSci2017;21:552–9.

[9] YuZ,LiT,HorngS-J,Pan Y, Wang H, Jing Y. An iterative locally auto-weighted least squares method for microarray missing value estimation. IEEETransNanobioscience2016;16:21–33.

[10] HronK,TemplM,FilzmoserP. Imputation of missing values for compo-sitionaldatausingclassicalandrobustmethods.ComputStatistDataAnal2010;54:3095–107.

[11] Ching WK, Li L, Tsing NK, Tai CW, Ng TW, Wong AS, Cheng KW. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. Int J Data Min Bioinform. 2010;4(3):331-47. doi: 10.1504/ijdmb.2010.033524. PMID: 20681483.

[12] JahanF,SinhaNC,RahmanMM,RahmanMM,MondalMSH,IslamMA.Com-parison of missing value estimation techniques in rainfall data of Bangladesh.TheorApplClimatol2019;136:1115–31.

[13] PedersenAB,MikkelsenEM,Cronin-FentonD,KristensenNR,PhamTM,PedersenL,PetersenI.Missingdataandmultipleimputationinclinicalepidemiologicalresearch.ClinicalEpidemiol2017;9:157.

[14] Song Q, Ge H, CaverleeJ, Hu X. Tensor completion algorithms in big dataanalytics.ACMTransKnowledgeDiscoveryData2019;13:1–48.

[15] LiuS,DaiH.Examinationofreliabilityof missing value recovery in datamining. In: 2014 IEEE international conference on data mining workshop, IEEE;2014.p.306–13.

[16] Chi EC, Zhou H, Chen GK, Del Vecchyo DO, Lange K. Genotype imputation viamatrixcompletion.GenomeRes2013;23:509–18.

[17] AzurMJ,StuartEA,FrangakisC,LeafPJ.Multipleimputationbychainedequations: what is it and how does it work? Int J Methods Psychiatric Res2011;20:40–9.

[18] ValdiviezoHC,VanAelstS.Tree-basedpredictiononincompletedatausingimputationorsurrogatedecisions.InformSci2015;311:163–81.

[19] Xu X, Xia L, Zhang Q, Wu S, Wu M, Liu H. The ability of different imputationmethods for missing values in mental measurement questionnaires. BMC MedResMethodol2020;20:1–9.

[20] KhotimahBK,SuprajitnoH,etal.Modelingnaïvebayes imputation classifi-cationformissingdata.In:IOPconferenceseries:Earthandenvironmentalscience.243,IOPPublishing;2019,012111.

[21] AudigierV,HussonF,JosseJ. Multiple imputation for continuous vari-ables using a Bayesian principal component analysis. J Stat Comput Simul2016;86:2140–56.

[22] JosseJ,PagèsJ,HussonF.Multipleimputationinprincipalcomponentanalysis.AdvDataAnalClassif2011;5:231–46.

[23] Tsai, Chih-Fong et al. “A class center based approach for missing value imputation.” *Knowl. Based Syst.* 151 (2018): 124-135.

[24] ChenX.Animprovedself-representationapproach for missing value imputation.In:201824thinternationalconferenceonpatternrecognition.IEEE;2018, p.1450–5.

[25] WeiR,WangJ, SuM,JiaE,ChenS,ChenT,NiY.Missingvalueimputationapproachformassspectrometry-basedmetabolomicsdata.SciRep2018;8:1–10.

[26] Arciniegas-AlarcónS,García-PeñaM,KrzanowskiW,dosSantosDiasCT.Imputing missing values in multi-environment trials using the singularvaluedecomposition:Anempiricalcomparison.CommunBiometryCropSci2014;9:54–70.

[27] ChoudhurySJ,PalNR.Imputationofmissingdatawithneuralnetworksforclassification.Knowl-BasedSyst2019;182:104838.

[28] Kaiser J. Algorithm for missing values imputation in categorical data with useofassociationrules.2012,ArXiv:1211.1799.

[29] PatilBM,JoshiRC,ToshniwalD.Missingvalueimputationbasedonk-meanclustering with weighted distance. In: International conference on contemporarycomputing,vol.94.Springer;2010,p.600–9.

[30] Feng X, Wu S, Liu Y. Imputing missing values for mixed numeric and categoricalattributesbasedonincompletedatahierarchicalclustering.In:Internationalconferenceon knowledgescience,engineeringandmanagement,vol.7091.Springer;2011,p.414–24.

[31] MigdadyH,Al-TalibMM.AnenhancedfuzzyK-meansclusteringwithapplicationtomissingdataimputation.ElectronJApplStatAnal2018;11:674–86.

[32] TangJ, ZhangG,WangY,WangH,LiuF.A hybrid approach to integratefuzzyC-meansbasedimputationmethodwithgeneticalgorithmformissingtrafficvolumedataestimation.TranspResC2015;51:29–40.

[33] SinghN,JaveedA,ChhabraS,KumarP.Missingvalueimputationwithunsupervised kohonenself organizing map. In: Emerging research in computing,information,communicationandapplications.Springer;2015,p.61–76.

[34] Loh W-Y, EltingeJ, Cho MJ, Li Y. Classification and regression trees and forestsforincompletedatafromsamplesurveys.StatistSinica2019;29:431–53.

[35] Minakshi, Vohra R, Gimpy. Missing value imputation in multi attribute dataset.IntJComputSciInfTechnol2014;5:1–7.

[36] AbdullahSS,MalekMA,AbdullahNS,KisiO,YapKS.Extremelearningmachines: a new approach for prediction of reference evapotranspiration. JHydrol2015;527:184–95.

[37] LobatoF,SalesC,AraujoI,TadaieskyV,DiasL,RamosL,SantanaA.Multi-objective genetic algorithm for missing data imputation. Pattern Recognit Lett2015;68:126–31.

[38] Huang J, Sun H. Grey relational analysis based k nearest neighbor missing dataimputation for software quality datasets. In: 2016 IEEE international conferenceonsoftwarequality,reliabilityandsecurity.IEEE;2016,p.86–91.

[39] Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Softw2012;85:2541–52.

[40] Yang F, Du J, Lang J, Lu W, Liu L, Jin C, Kang Q. Missing value estimationmethodsresearchforarrhythmiaclassificationusingthemodifiedkerneldifference-weightedKNNalgorithms.BioMedResInt2020;2020.

[41] ZhuX,ZhangS,JinZ,ZhangZ,XuZ. Missing valueestimation formixed-attributedatasets.IEEETransKnowlDataEng2010;23:110–21.

[42] Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L.Missing data imputation using statistical and machine learning methods in arealbreastcancerproblem.ArtifIntellMed2010;50:105–15.

[43] KoklaM,VirtanenJ,KolehmainenM,PaananenJ,HanhinevaK.Randomforest-based imputation outperforms other methods for imputing LC-MS metabolomicsdata:acomparativestudy.BMCBioinformatics2019;20:1–11.

[44] TangJ,ZhangX,YinW,ZouY,WangY.Missingdataimputationfortrafficflow based on combination of fuzzy neural network and rough set theory. JIntellTranspSyst2021;25:439–54.

[45] Wang L, Fu D, Li Q, Mu Z. Modelling method with missing values based onclusteringandsupportvectorregression.JSystEngiElectron2010;21:142–7.

[46] MyersTA.Goodbye,listwisedeletion:Presentinghot deck imputation as aneasy and effective tool for handling missing data. Commun Methods Measures2011;5:297–310.

[47] Pati SK, Das AK. Missing value estimation for microarray data through clusteranalysis.KnowlInfSyst2017;52:709–50.

[48] Aussem A, de Morais SR. A conservative feature subset selection algorithm withmissingdata.Neurocomputing2010;73:585–90.

[49] Ghannad-RezaieM,Soltanian-ZadehH,YingH,DongM.Selection–fusionapproachforclassificationofdatasetswithmissingvalues.PatternRecognition2010;43:2340–50.

[50] Brahma, B., Kamila, N. K., Dhal, S. K., Pani, S. K., Mahesh, N., & Majhi, S. K. (2021, May 12). An extensive evolutional survey of Medical Domain Data Analytics & Decision Improvisation Systems. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3842573

[51] Brahma, B., Bhuyan, H.K. (2022). Soft Computing and Machine Learning Techniques for e-Health Data Analytics. In: Mishra, S., González-Briones, A., Bhoi, A.K., Mallick, P.K., Corchado, J.M. (eds) Connected e-Health. Studies in Computational Intelligence, vol 1021. Springer, Cham. https://doi.org/10.1007/978-3-030-97929-4_4

[52] Brahma, B., et. al. (2021). Mathematical Model forAnalysis of COVID-19 Outbreak using VON Bertalanffy Growth Function (VBGF). Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 6063-6075. https://doi.org/10.17762/turcomat.v12i11.6925

[53] Baitharu, T.R.; Pani, S.K. Effect of Missing Values on Data Classification. J. Emerg. Trends Eng. Appl. Sci. (JETEAS) 2013, 4, 311–316.

[54] Mishra AK, Pani SK, Ratha BK (2015) Association rule mining with apiori and FPGrowth using Weka.Int J Adv Technol Eng Sci 3(1)

[55] G. Panda, S.K. Dhal, R. Satpathy, S.K. Pani(2022) ANFIS for fraud automobile insurance detection system Advances in data science and management, Springer, Bhubaneswar, India , pp. 519-530.

[56]