Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

A Survey On Missing Data in Machine Learning

Tlamelo Emmanuel (tlameloemmanuel@studentmail.biust.ac.bw)
Botswana International University of Science and Technology https://orcid.org/0000-0002-6340-063X

Thabiso Maupong
Botswana International University of Science and Technology

Dimane Mpoeleng
Botswana International University of Science and Technology

Thabo Semong
Botswana International University of Science and Technology

Mphago Banyatsang
Botswana International University of Science and Technology

Oteng Tabona
Botswana International University of Science and Technology

Survey paper

Keywords: Missing Data, Imputation, Machine Learning

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-535520/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

A survey on missing data in Machine Learning

Tlamelo Emmanuel*, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Mphago Banyatsang and Oteng Tabona

*Correspondence: tlamelo.emmanuel@studentmail.biust.a
Department of Computer Science and Information System, Botswana International University of Science and Technology, Palapye, Botswana
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Machine learning has been the corner stone in analysing and extracting information from data and often a problem of missing values is encountered. Missing values occur as a result of various factors like missing completely at random, missing at random or missing not at random. All these may be as a result of system malfunction during data collection or human error during data pre-processing. Nevertheless, it is important to deal with missing values before analysing data since ignoring or omitting missing values may result in biased or misinformed analysis. In literature there have been several proposals for handling missing values. In this paper we aggregate some of the literature on missing data particularly focusing on machine learning techniques. We also give insight on how the machine learning approaches work by highlighting the key features of the proposed techniques, how they perform, their limitations and the kind of data they are most suitable for. Finally, we experiment on the K nearest neighbor and random forest imputation techniques on novel power plant induced fan data and offer some possible future research direction.

Keywords: Missing Data; Imputation; Machine Learning

Introduction

Missing values or data is a prominent area when dealing with data analysis. These missing values are usually attributed to: human error when processing data, machine error due to the malfunctioning of equipment, respondents refusal to answer certain questions, drop-out in studies and merging unrelated data [1, 2]. The best possible technique to handle missing values is to attempt to avoid the problem with careful planning, data collection and preparation. Despite vast efforts to avoid the problem, some missing values are common and unavoidable [3]. The problem is usually present in all domains that deal with data, be it machine learning, statistics or data-driven control and it causes different issues. These issues are performance degradation, data analysis problems and biased outcomes lead by the differences in missing and complete values [4]. Moreover, the seriousness of these missing values depend in part on how much data is missing, the pattern of missing data, and the mechanism underlying the missingness of the data [5]. The missing values can be handled by certain techniques including, deletion of instances and replacement with potential or estimated values [6–8], a technique denoted as imputation [9]. The imputation technique is regarded as a more complicated solution that considers several factors to try to handle missing values [10]. This is because it may lead to computation and analysis

irregularities and may also cause systematic differences in the data. Deletion of instances with missing values is an easy way out and somewhat a straightforward approach, but the end results would be the loss of valuable data. Apart from deletion, imputation is the most adopted approach for handling missing values especially on the pre-processing step in data analysis. Several statistical and machine learning techniques such as mean imputation, regression, K nearest neighbor, ensemble based imputation etc, have been proposed in the literature using this approach [11, 12]. However, machine learning methods have proven to be more effective in handling missing values in recent years compared to statistical techniques. Machine learning techniques are separated into two classifications, being supervised and unsupervised learning and have been extensively utilized in missing data, more especially the supervised learning technique. The supervised learning approach mostly relies on labelled data and ignore the possible role of unlabeled data to predicts the missing values [13]. While the other hand, the unsupervised algorithm learns on its own, from the unlabelled data to extract features and patterns for missing data [14]. In some cases, hybrid approaches [15–19], have been utilized to solve the weaknesses of the traditional supervised and unsupervised imputation methods. However, it is important to note that the only suitable solution comes down to a virtuous design and good analysis [20]. This is because analysis of performance is dependent but not limited to several factors such as the type of algorithm selected, attribute selection and sampling techniques. Also, as the era of big data is here, data has become large and complex that it is difficult to deal with using traditional learning methods since the established process of learning from conventional datasets was not designed to and does not work well with big data [21]. Therefore, when dealing with missing data, approach is always crucial since improper handling may lead to drawing inaccurate inferences.

In this study, we discuss missing values in section Missing Data Patterns and Mechanisms, where we also introduce missing data patterns and mechanisms. Section Missing Values Approaches empirically discusses approaches in the literature for handling missing values and critically review several implementations in different domains, mostly focusing more on machine learning. In section Performance metrics for missing data imputation, we discuss several performance metrics in the missing values domain and section Comparisons we discuss and analyse results from previous works. We then implement two machine learning algorithms using the Iris data-set on section Experimental evaluation on Machine Learning Methods and discussed the results. Finally, section Conclusion and Future Work summarises the paper and point out potential directions for future exploration.

Missing Data Patterns and Mechanisms

In this section, we discuss the missing patterns in data and different missing data mechanisms.

Missing Data Patterns

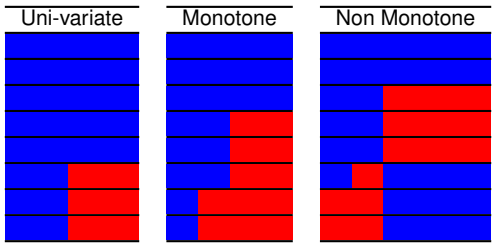
Missing data patterns describe which values are missing and observed in a data set. However, there is no standard list of missing data patterns in the literature as discussed in [22–24]. In this subsection, we discuss three missing data patterns that appear most in the literature which are uni-variate, monotone and non-monotone. In Table 1 we further demonstrates the different patterns in missing data.

Uni-variate: Missing data pattern is uni-variate when there is only one variable with missing data [25]. This pattern is rare in most disciplines and arises in experimental studies [26].

Monotone: Missing data pattern is said to be monotone if the variables in the data can be arranged, this pattern is usually associated with a longitudinal studies where members drop out and never return [27]. The monotone data pattern is easier to deal with since patterns among the missing values are easily observable [28].

Non Monotone: This is a missing data pattern whereby the missingness of one variable does not affect the missingness of any other variables[29].

Table 1: Representation of missing data patterns data. Blue represents observed values, red is missing values



Missing Data Mechanisms

Mostly mechanisms that lead to the missing values on data affect some assumptions supporting most missing data handling methods, hence, in the literature the missing data has been defined according to these mechanisms. The authors of [30] established the missing data theory, categorized by three main mechanisms for missingness, which are defined depending on the available and missing data. To define missingness, let Y be a matrix of the entire data set that is decomposed into Y_o and Y_m, which denote the observed and missing data. Let R denote a missing value matrix defined by,

$$R := \begin{cases} 0 \text{ , if Y is observed} \\ \\ 1 \text{ , if Y is missing} \end{cases}$$

Let q represent a vector of values that indicate the association between missingness in R and the data set Y. The missing values mechanisms are therefore defined by the probability of whether a value is observed or missing as we outline below.

0.0.1 Missing Completely at Random (MCAR)

This is when missing observations are not reliant on the observed and unobserved measurements. The probability of MCAR is defined as:

$$p(R|q) \tag{1}$$

Missing at Random (MAR)

The likelihood of a missing value in MAR is only related to the observable data. The probability for MAR can be defined as:

$$p(R|Y_o, q) \tag{2}$$

Missing at random (MAR) is mostly encountered in health science studies data sets. Under this mechanism, missing values can be handled by observed predictor variables[31].

0.0.2 Missing Not at Random (MNAR)

This refers to when missing data is neither MCAR nor MAR. The missing data depends equally on the missing and observed values. In this method, handling the missing values is usually impossible, as it depends on the unseen data. The MNAR probability is defined as:

$$p(R|Y_o, Y_m, q) \tag{3}$$

The probability of whether a position R is missing or observed depends on both Y_o and Y_m.This mechanism is mostly applied in different domains predominantly in the domain of (bio)medicine [32], but is also applied in the psychological and educational data-sets [33, 34].

According to [11, 35], it is mostly impossible to unambiguously categorise missing data into these three mechanisms since imagining that missing data is completely not related to other non missing variables is very challenging because one way or the other missing values relate to non-missing variables. Many researchers, however, report that the easiest way is to complete all the missing data as MAR to some degree because MAR resides in the middle of this continuum [11].

Missing Values Approaches

In this section we discuss missing values approaches available in the literature. We also review implementation of missing values approaches in various domains.

Deletion

In this approach all entries with missing values are removed/ discarded when doing analysis. Deletion is consider the most simple approach as there is no need to try and estimate value. However, the authors of [22] have demonstrated some of the weakness of deletion, as it introduce bias in analysis, especially when the missing data is not randomly distributed. The process of deletion can be carried out in two ways, pairwise or list-wise deletion [3].

List-wise or case deletion

In list-wise deletion, every case that has one or more missing values is removed. List-wise deletion has become the default choice when analysing data in most statistical software packages [36]. However, under the assumption that the data is not MCAR, list-wise results in biasness [37]. While, if the data samples are large enough and the MCAR assumption is satisfied, then list-wise deletion may be a reasonable approach. In the event that the sampled data is not large, or the MCAR assumption is not satisfied, then list-wise deletion is not the best approach to consider. List-wise deletion may also result in losing so important information, especially when the discarded cases are high in numbers.

Pairwise deletion

In order to mitigate against information loss when doing do list- wise deletion one can use pairwise deletion. This is because pairwise deletion is carried out such that it reduces losses that could occur in list-wise deletion. This is done by eliminating values only when there is a certain data point needed to test if the value assumed to be missing is in fact missing [38]. The weakness of pairwise deletion is that it can lead to an inter-correlation matrix that is not positive definite, which is can possibly prevent further analysis such as calculating coefficients estimates [39]. Finally, pairwise deletion also known to produce low bias results for MCAR or MAR data [37].

Imputation

The process of imputation involves replacing missing values by some predicted values. The non-missing values data set is normally used to predict the values used to replace the missing values [9]. In the following we cover some of the most used imputation methods in the literature.

Simple imputation

Simple imputation approach entails replacing missing values for each individual value by using a quantitative attribute or qualitative attribute of all the non-missing values [40]. With simple imputation, missing data is handled by different methods such as, mode, mean, or median of the available values. In most studies simple imputation methods are used because of their simplicity and that they can be used as an easy reference technique [41]. However, simple imputation methods may produce bias or unrealistic results on a high-dimensional data sets. Also, with the generation of big data emerging, this method seems to be performing poorly and therefore is inadequate to be implemented on such data sets [42].

Regression imputation

Regression is one of the preferred statistical technique for handling missing values. This method is also termed conditional mean imputation, here missing values are replaced with a predicted value created on a regression model if data is missing at random. The overall regression process is a two-phase method: the first step, uses all the complete observations to build a regression model, and imputes missing data based on the built regression model [43]. The regression method is decent since it maintains the sample size by preserving all the observations with missing values. However, regression may need a large sample of data to produce stable results. Furthermore, a single regression curve is followed for all the imputed values and no inherent variation is presented in the data [22]. Considering a feature containing missing values, and the remaining attributes are complete. A regression model approximates the missing features using the available data. The first step is to estimate a set of regression equations that will predict the incomplete values from the complete values using a complete case. Predicted values are then generated for the incomplete variables. These predicted values fill in the missing values.For the imputation of y variables given a set of variables $j_1, ..., jq$, a regression model is used as follows:

$$y = \alpha + \beta_1 j_1 + ... + \beta_q j_q + \epsilon \tag{4}$$

With α, $\beta_1, ..., \beta_q$ being the unknown values and ϵ is a distance variable. The estimates in equation 4 will results in a prediction for y given by the variables:

$$\hat{y} = a + b_1 j_1 + ... + b_q j_q \tag{5}$$

with a b_1 b_q denoting the least squares estimates of α, $\beta_1, ..., \beta_q$. An imputation $\tilde{y}$ is then made

$$\tilde{y} = \hat{y} = a + b_1 j_{1_i} + ... + b_q j_{q_i} \tag{6}$$

The technique of regression implemented depend on the nature of the data. If there are two or more missing features, a multivariate regression model has to be used for imputation [44]. Multivariate Regression measures the degree at which more than one independent prediction and more than one dependent responses, are linearly related [45]. A multivariate regression imputation is used as follows using the extension of a standard regression model in equation 4:

$$y = \mu_y + B_{yj}(j - \mu_j) + \epsilon \tag{7}$$

where the target value in y is retrieved by using the same vector of variables j. An expectation maximization algorithm is then used to find the estimates of the parameters in 7, the algorithm uses the information of the observed data to estimate the parameters. More information on the expected maximisation is presented on [46]. After obtaining estimates of the unknown parameters in equation 7, the imputation of missing values in y is obtained as before from the observed vector j_i. Then an imputation is retrieved directly from the predicted value,

$$\tilde{y_i} = \hat{y_i} = \hat{\mu}_y + \hat{B_{yj}}(j_i - \hat{\mu}_j) \tag{8}$$

and an imputation is done by adding a random disturbance to the prediction:

$$\tilde{y_i} = \hat{y_i} + e_i = \hat{\mu}_y + \hat{B_{yj}}(j_i - \hat{\mu}_j) + e_i \tag{9}$$

A common choice is to get e_i from a multivariate distribution with a mean vector zero and the residual of the regressions y on j [46].

In research studies using the regression approach includes one by [47], where a weighted quantile regression approach that estimated missing values in health data was conducted. The authors used a quantile regression approach on the health data because it is usually attributed to a high level of skewness, heteroscedastic variances and the weighted quantile regression estimator is consistent,unlike the naive estimator, and asymptotically normal making it suitable for analysing this type of data. The experiment demonstrated the effectiveness of the quantile regression technique on the numeric health care cost data analysis. However, the estimator used fully observed observations and was most suitable when the rate of the missing data was not excessively high. Moreover, the approach was not robust due to functional form specification and could have introduced bias results.

In another study, the authors proposed a complete case regression missing values handling method using functional principal component [48]. The performance of the

approach when the missing values were not handled was experimented on and compared with regression imputed missing values. Their major interest in the study was the functional linear regression when some observations of the actual response were missing.

Another study applied a multivariate imputation technique for imputing missing values in normal multivariate data. The imputation values were obtained from the sequence of regression, where all the variables containing missing values were regressed against the variables that did not contain missing values as predictor variables by using the iteration approach. The approach worked well with more one variable containing missing values and non-monotonous patterns [49].

Hot-Deck Imputation

Hot-deck imputation handles missing values by matching the missing values with other values in the data set on several other key variables that have complete values [50]. The method has variations, but one that allows natural variability in missing data selects a pool of all cases. This pool is called the donor pool, that is identical to the cases with missing data on many variables and chooses one case randomly out of that pool. The missing value is then replaced by data from the randomly chosen cases. Another technique involves replacing the closest donor neighbor rather than selecting one donor from a pool of donors [51]. However, the method disregards the variability in missing data. The other variations of this imputation technique are weighted random hot-deck and weighted sequential hot deck. The weighted random hot deck method does not limit the number of times a donor is nominated, however, the donors are chosen randomly from the donor pool. In contrast, weighted sequential hot-deck puts a restriction on the amount of time a donor can be chosen to prevent the same donor to be paired with a large quantity of recipients [50].

The hot-deck method is very popular in all single imputation methods as it results in a rectangular data [50], that can be used by secondary data analysts. Also, the method avoids cross-user inconsistency and does not depend on model fitting for the missing value to be replaced, making it possibly less delicate to model specification as compared to a method built on a parametric model, for instance regression imputation. The method also decreases bias in non-response. Even though the method is being used widely in research, its concept is not as well established compared to other imputation techniques.

In [52], a hot deck imputation method that allowed for the investigation of the impact of missingness mechanisms, ranging from MAR to MNAR, and used the information contained in fully observed covariates was proposed. Bias and coverage of estimates from the proposed technique were investigated by simulation. Results also, showed that the method performed best when fully observed values were associated with the outcome.

In another study [53], a fractional hot deck imputation method was used to handle missing values. The procedure was applied to the MAR mechanism, but the missing data pattern and the comparison was done with list-wise deletion, mean, median imputation methods only. Their method produced a smaller standard error compared to other method they used for comparison. However, the experiment may have been bias since it was concluded that it performed better being compared to the imputation method that usually produce biased results.

Expectation-maximization

The expectation maximization technique is an iterative method for handling missing values in numerical datasets, the algorithm uses an *"impute, estimate and iterate until*

convergence" approach. Every iteration includes two stages which are: expectation and maximisation. Expectation estimates missing values given observed data, whereas in maximisation, the present estimated values are used to maximize the probability of all the data [54].

Approaches in research have been proposed to deal with missing values using expectation minimisation. In [55], an investigation on handling missing data was done using a dataset that analysed the impacts of feeding behaviors among drug-treated and untreated animals. The expectation maximisation algorithm was used and compared to other methods like list-wise deletion which was the least efficacious method, Bayesian approach and the mean substitute regression. The authors concluded that that the EM algorithm was the best method for the type of data they used. However, using real datasets in the study may have led to the results being specific to idiosyncrasies in the dataset and in sampling or are reflective of hypothetical expectations.

In another research, an expected maximisation algorithm was used for imputation to solve the problem of training Gaussian mixtures in large high-dimensional datasets settings with missing values [56]. The imputed datasets were then experimented in classification models and proved to provide a significant performance improvement over other basic missing value imputation methods. However, the technique resulted in expensive matrix computations.

Generally, single imputation methods as discussed above are simple methods to handle missing data and save time. However, these methods are mostly bias, and error of their imputations is not incorporated. Furthermore, single imputation techniques do not represent the vulnerability associated with the missing values [11]. Therefore, researchers have experimented on improved methods to handle missing data that give much better performance [12]. The improved techniques are believed to be unrivaled to the single missing data techniques since they proved to yield unbiased analysis.

Multiple Imputation

It is evident that missing data handling goes beyond deleting or discarding missing data [30] and therefore researchers resort to multiple imputation. Multiple imputation is where the distribution of the observed data is utilized to approximate numerous values that reflect the uncertainty around the true value, and this method was mostly implemented to solve the limitations of single imputation [57]. The analysis is done on a data set using the various missing data techniques, and the average of parameter estimates across M samples is computed into a single point estimate. Thus, multiple imputation technique comprises of three distinct phases:

- Missing data is handled in M resulting in M complete data sets.
- The M complete data sets are then analysed.
- The results of all the M imputed data sets are combined for the final imputation result.

Though multiple imputation is set up as a standard methodology for dealing with missing values, it is important for researchers to utilize appropriate techniques for imputation, to guarantee that dependable results are obtained when experimenting with this approach [58]. Furthermore, performance may be affected negatively when carrying out imputation on real data such as survey data, clinical data and industrial data which may be characterized by a high rate of missingness and a great number of factors that are not necessarily linearly

related. Also, traditional multiple imputation methods seem to perform poorly on high dimensional data and researchers have resorted to improving these algorithms to enhance their performance [59, 60].Similarly, there is also evidence that caution should be made for continuous-based techniques when imputing categorical data as this may lead to biased results [61].

We discuss the approaches on the literature on multiple imputation: The researchers in [61], experimented on a technique that accurately imputed missing values on a patient data set using multiple imputation using Least Squares Support Vector Machine(LSSVM). Five datasets were used to determine the performance of the proposed method. The evaluation results illustrated that their method outperformed conventional imputation methods and that the study was a more robust technique that generated values closer to the one that was missing. Moreover, the author also proposed another method Clustered Z-score Least Square Support Vector Machine(CZLSSVM) and demonstrated its efficiency in two classification problems for incomplete data. Their experimental results also indicated that the accuracy of the classification was increased with CZLSSVM and that the algorithm outperformed other data imputation approaches like SVM, decision tree, KNN, rough sets and artificial neural networks. In another study [62], the authors also proposed a multiple imputation method for clinical practice data. The results of the method gave unbiased estimates and standard errors, on MCAR or MAR missing mechanisms. Also, the prediction model specification was adequate, though it may have required the help of a statistician. However, their multiple imputation technique performed better than the other conventional methods. There has been a study also by [42], that explored a multiple imputation approach that extended multivariate imputation by chained equation for big data. The approach had presented two variants one for categorical and the other numeric data and implemented twelve existing algorithms for performance comparison. The experimental results of the experiment with four datasets demonstrated that the method performed better for the imputation of binary and numeric data.

Imputation methods inspired by machine learning

Imputation methods built on machine learning are sophisticated techniques that mostly involve developing a predictive approach to handle missing values using unsupervised or supervised learning. As other imputation methods these techniques estimate the missing data estimation depending on the information available from the non -missing values in the data using labelled or unlabelled data. Mostly if the available data has useful information for handling the missing values, an imputation high predictive precision can be maintained. We discuss some of the most researched on machine learning imputation techniques below.

K Nearest Neighbour classification

The KNN algorithm works by classifying the nearest neighbours of missing values and use those neighbours for imputation using a distance measure between instances [63]. Several distance measures such as the Minkowski distance, Manhattan Distance, Cosine Distance, Jaccard Distance, Hamming Distance and Euclidean distance can be used for KNN imputation, however the Euclidean distance is reported to give efficiency and productivity

[64], [65] and therefore is the most widely used distance measure. We further explain the KNN imputation using the Euclidean distance measure below:

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m}(X_{ik} - X_{jk})^2} \tag{10}$$

Where: $Dist_{xy}$: Is the Euclidian distance ,k is data attributes $j = 1, 2, 3.....k$, k data dimensions, (X_{ik}) : value for $j-$ attribute containing missing data and (X_{jk}) is the value of $j-$ attribute containing complete data.

The value of the k points that have a minimum distance are chosen then Weight Mean Estimation is calculated.

$$X_k = \frac{\sum_{j=1}^{J} w_j v_j}{\sum_{j=1}^{J} w_j} \tag{11}$$

Where: X_k is the mean estimation, J is the number of parameters used with j=1,2,3....K. v_j are complete values on attributes containing missing data while w_j is the nearest neighbors observed which the equation is:

$$X_k = \frac{1}{disc_{(x,y)^2}} \tag{12}$$

The KNN imputation technique is flexible in both discrete and continuous data and can also be implemented as a multiple missing data handler[1, 63]. However, KNN imputation has drawbacks such as low precision when imputing variables and introduces false associations where they do not exist [66]. The other weakness of KNN imputation is that it searches through all the data set, hence increasing computational time [67]. However, there are approaches in literature that have been developed to improve the KNN imputation algorithm for missing values problems, see[68–73].

A KNN imputation using several cases with different mechanisms and missing data models was proposed [74]. The authors concluded that their method performed well in handling missing values. However, the research did not follow any missing value mechanism when manually removing the data for the experiment, which may lead to bias results.

In another research, the authors introduced an iterative KNN imputation method which was an instance-based technique that took advantage of the correlation on the attributes by using grey relational grade as an alternative for Euclidean distance measure to search k-nearest neighbour instances. The imputed data was predicted from these nearest neighbour instances iteratively. This iterative imputation permitted all values from the preceding iteration to be used for missing value estimation. Also, the method was reported to fill in all the missing values with dependable data regardless of the missing rate of the dataset. The experimental results suggested that the proposed method resulted in a better performance than other methods regarding imputation accuracy and convergence speed [75]. However, the dataset that was used here had originally no missing values and the

missing values been imputed at random not considering other missing values mechanisms which may have led to unrealistic results.

In another research, a novel grey relational analysis approach for incomplete instances using the KNN imputation technique was experimented on[76]. The approach was experimented on four datasets with different artificial missingness set-ups to investigate the performance of the imputation. The experiential results of the study demonstrated that the approach was superior to traditional KNN imputation. Furthermore, the classification accuracy could be maintained or improved by using this approach in classification tasks.

Another study developed a novel K Nearest Neighbour (KNN) incomplete-instance based imputation approach called CVBkNN, which utilized cross-validation to improve the parameters for each missing value [77]. Eight different datasets were used for the experiment. The results of the study demonstrated that their approach was superior to other missing values approaches. They also displayed the optimal fixed parameter settings for KNN imputation for software quality data. Their approach proved to improve classification accuracy or at least maintained it. However, determining additional meaningful parameters for configuration could have improved the study's accuracy further.

In another study by [78], the KNN algorithm was experimented to evaluate its efficiency as an imputation method to treat missing data and compared its performance to other algorithms such as by the C4.5 and CN2 and the mean or mode imputation method. In the experiment missing values were artificially implanted, in different rates and attributes, into the data sets. The KNN algorithm performed well even in the presence large amount of missing data compared to the other algorithms.

A genetic algorithm enhanced k- nearest neighbour for handling missing values named EvlKNNImpute was also proposed in this study. The KNNImpute has showed effective compared to other methods used in imputation using the yeast dataset [79]. Their approach also proved to perform better when there was an elevated level of missing rate in a data than a small missing rate.

In another study, the authors incorporated correlation matrix for KNN algorithm design. The least-squares loss function was used to minimize the reconstruction error and reconstruct every test data point by using all training data points. Their method, compared with traditional KNN methods, proved to achieve a higher accuracy and efficiency [80]. However, like many other kinds of research in data imputation this study did not consider the influence of missingness mechanisms and patterns on imputation performance.

The KNN imputation method has been highly researched for imputation since it has proved in literature to perform better than other imputation approaches as seen in the reviews above. However, none of the studies systematically analysed the effects of imputation ordering in the KNN imputation performance. Moreover, there is still no proven common resolution to select the optimized KNN parameters for imputation. Although some researchers use different missingness scenarios to evaluate their approaches, the significance of the influences of this missingness mechanisms are often neglected. Also, the use of KNN in the big data setting is still an under-explored area.

Support Vector Machine (SVM)

Another common machine learning algorithm that is extensively used for missing data handling is the SVM [81], [82]. The SVM, for a labelled training sample, efforts to discover

an optimal separating hyper-plane such that the distance from the hyper-plane to the nearest data points is maximized [83]. The hyper-planes are defined by

$$w \cdot x_1 + b \geq +1 \ when \ y_i = +1 \tag{13}$$

$$w \cdot x_1 + b \leq -1 \ when \ y_i = -1 \tag{14}$$

Where w is a weight vector, x is an input vector and b is bias.

Like other machine learning algorithms, the imputation of missing values with this method can impact the accuracy and utility of the resulting analysis. Authors of [81], used the SVM regression-based method for missing data imputation. A decision value was set as a condition value and the condition value as the decision value and the SVM regression was used to predict the condition values. The experimental results proved that the SVM regression approach had the highest precision on the SARS data set. However, the experiment did not report any use of missing value patterns, ratios or mechanisms used. Also, in [84], the authors demonstrated an SVM and Gaussian processes for missing data handling using exponential families in feature space. In this research estimation with missing values become a problem of computing marginal distribution and finding efficient optimization methods. In another approach [85], the authors replaced the missing values by using the results obtained from applying the SVM classifier over the training set and also used an SVM regression to hanlde the values. The authors experimented using the SVM classifier as an imputation approach because it was reported to perform well on text categorisation problems in [86].However the results of the study concluded that the SVM regression approach gave a much better performance compared to the SVM classifier and other classification and regression approaches, though this might have been influenced by the imbalanced dataset used for the experiment. Since imbalanced data may contribute to the increase of performance of SVM regression.

In [87], handled missing values by max-margin learning framework. They formulated an objective function, which used geometric interpretation of the margin, that aimed to maximize the margin of every sample in its own relevant subspace. They also showed two approaches for optimizing the general case: an estimation that can be solved as a standard quadratic problem and an iterative approach for solving the exact problem. Their methods saved computational time by avoiding the pre-processing step. More importantly, they demonstrated an elegant missing value handling approach which outperformed other methods when the missing values had a significant structure, and the approach also proved to be competitive compared with other techniques when the values are missing at random.

Decision Tree

The decision tree is a machine learning algorithm that illustrates all conceivable outcomes and the paths leading to those outcomes in the form of a tree structure. Missing values imputation using this method is done by building decision trees to observe the missing values of each variable, and then fills the missing values of each missing variable by using its corresponding tree [88]. The missing values prediction is then shown in the lead node. Additionally, this algorithm can handle both numerical and categorical variables, identify

the most variables and eliminate the rest. However, decision trees can produce a complex trees that tend to be time consuming, but have a low bias [89].

Several researchers[85, 90–92] have used decision trees for imputation, and we discuss their input. A decision tree and forest technique for the imputation of categorical and numerical missing values was proposed. The technique identified horizontal segments in the data set where the records belonging to a certain segment had higher similarity and attribute correlations. The missing data were then imputed using the similarity and correlations. Nine real life data sets were used to compare the technique to other existing ones using four regularly used evaluation criteria [90]. Their experimental results indicated a clear superiority of the technique. However, an improvement on their technique for attaining a better computational complexity, and memory usage may be needed.

Also, in a by [91], a missing values approach using a decision tree algorithm. A student data set with missing values was used and a classification algorithm was implemented for comparing accuracy with incomplete data and after imputation. As a result, accuracy was higher on imputed data set as compared to incomplete data set. However, in this study there was no report on missingness ratios or mechanisms considered.

In another paper [92], the authors presented a missing value handling technique, using decision trees and expectation-maximization algorithm. They argued that the correlations among the attributes in the horizontal partition of a data set could be higher than the correlations over the whole data set. Also, that expectation maximization performance on higher correlations data is expected to be better than on lower correlations data set. Therefore, they applied expected maximization imputation on various horizontal segments of the data with high correlations between the attributes. Also, various patterns of missing values with different missing ratios were used and the experimental results indicated that their approach performed significantly better.

Another study replaced the missing values by applying the Decision Trees approach. The authors pruned the decision tree by learning the pruning confidence over a training set and predicted probabilities keeping the minimum number of instances per leaf to 2. The method was proposed with other methods for handling missing data and the author concluded that the results of different approaches were dataset dependent and no approach was a solution for all [85].

The three most used decision tree learning algorithms are: *ID3, C4.5 and CART*.

- CART: Classification and Regression Trees (CART) addresses both continuous and categorical values to generate a decision tree and handle missing values.The algorithm identifies a two fold rule based on one indicator variable that segments the data into two nodes by minimizing variance of the outcome within each node. The tree is then developed by proceeding this splitting recursively until reaching a stopping point determined by the tuning parameters. Imputation is then made from a regression tree by identifying the terminal node to which a new subject belongs and sampling from the outcomes in that node [93]. An attribute selection measure Gini Indexing is used in CART to build a decision tree which unlike ID3, C4.5 does not use probabilistic assumptions. Also, CART generates binary splits that produce binary trees which other decision tree methods do not. Furthermore, this method uses cost complexity pruning to remove the unreliable branches from the decision tree to improve accuracy and does not rely upon distributional assumptions on the data [94].

- ID3: This is a decision tree technique that can be built in two stages: tree building and pruning. A top-down, greedy search is applied through a given set to test each attribute at every tree node. Then information gain measure is used to select the splitting attribute. It only accepts categorical attributes when building a tree model and does not give precise outcome when there is noise. Continuous missing values can be handled by this method by discrediting or considering the value for the best split point and taking a threshold on the attribute values. This method does not support pruning by default, however, it can be done after building a data model [94].
- C4.5: This algorithm was developed after the ID3 algorithm and handles both continuous and categorical values when constructing a decision tree. C4.5 addresses continuous attributes by separating the attribute values into two portions based on the selected threshold such that all the values above the threshold is regarded as one child and the remaining as another child. Gain Ratio is used as an attribute selection measure to construct a decision tree. The algorithm handles missing values by selecting an attribute using all instances of a known value for information gain calculation. Instances with non missing attributes are then split as per actual values and instances with missing attribute are split proportionate to the split off known values. A test instance with missing value is then split into branches according to the portions of training examples into all the child nodes [95]. The algorithm withdraws bias information gain when there are many output values of an attribute.

Another popular form of the Decision trees approach is the Random Forest algorithm, which is a stack of decision trees through bagging which combines multiple random predictors in order to aggregate predictions the prediction rule is based on the majority vote or average over all trees. Forests are able to achieve competitive or even superior prediction strengths in comparison to well established approaches such as regression and support vector machines [96]. The process of imputing missing values with the random forest include [97]:

1. Selecting a random sample of the observations with replacement;
2. A set of variables are then selected at random;
3. A variable providing the best split is chosen;
4. The step of choosing a variable that produces the best split is repeated until the maximum depth is reached;
5. The steps above are repeated until the certain number of trees is reached;
6. A prediction of the missing value is then done upon a majority vote.

There are several studies in literature [98, 99], where Random Forests were used for handling missing values. In a study by [100] an extensive simulation study that involved missing at random simulated datasets using random forest imputation and evaluated in comparison with predictive mean matching.

Clustering Imputation

Clustering methods, such as hierarchical clustering and k-means clustering have been generally experimented for missing data handling in the literature. The K-means clustering technique consists of 2 steps where, in the first step K-means clustering is used to get clusters, then the cluster information is used to handle the missing values [101]. However, clustering methods are reported to not be robust enough to handle the missing data problem. The clustering method can be defined as follows [102]:

Given a data set $T = t_1, t_2, ...tp, ...T_{N_p}$ where T_p is a feature vector in the N_d- dimensional feature space,this feature vector t is a single data point and N_p is the number of patterns in T, then the clustering of T is the partitioning of T into K clusters $C_1, C_2, \ldots, C_K$ satisfying the following conditions:

- Every feature vector has to be assigned to a cluster

$$\bigcup_{k=1}^{K} C_k = T \tag{15}$$

- With at least one feature vector assigned to it

$$C_k \neq \phi, k = 1, ..., K \tag{16}$$

- Each feature vector is assigned to one and only one cluster

$$C_k \bigcap C_{kk} = \phi \tag{17}$$

where $k \neq kk$

In study by [101], a missing value imputation method was proposed based on K-means clustering . The proposed method was applied to clinical datasets from the UCI Machine Learning Repository. The method proved to perform better than the simple method that did not use imputed values for further imputations. However, errors in earlier imputations may have propagated to further imputations. Hence this point should be considered when applying methods like the proposed method on real world datasets. In another paper, a clustering-based non-parametric kernel-based imputation technique, called Clustering-based Missing value Imputation (CMI), was proposed for dealing with missing values in target attributes [103]. The experimental results demonstrated the algorithm was an effective method in creating inference for variance and distribution functions after clustering. However, the approach did not consider missing values in conditional attributes and class attributes. There has also been advances in imputing big data based on clustering, [104] proposed a big data k-means clustering, and a big data fuzzy k-means missing values approach that resulted in robust and efficient output for big data and offered reasonable execution times. The two imputation techniques surpassed in most cases mean imputation and elimination of the instances with lost values during classification. offer robust and efficient results for Big Data datasets, offering reasonable execution times. The fuzzy k-means approach was proved to provide better results for high percentages of missing values in the data, while the k-means performed better with the dataset that had lower amounts of missing values. Zhang et al [105], also proposed a multiple imputation clustering based approach that handled missing values in big longitudinal trial data in e-Health. The proposed concept proved that it could be easily adapted for different types of clustering for big incomplete longitudinal trial data in eHealth services.

Ensemble Methods

Ensemble methods are strategies that make multiple models and then combine them to produce a single improved result. This methods usually produces more precise results than a single model would. This has been the case in a number of machine learning competitions,

where the triumphant models used ensemble techniques [106]. Studies have confirmed that ensemble missing data handling algorithms outperform single base machine learning algorithms [107–111]. Also, ensemble methods can be implemented in parallel computing environments, which are necessary to process missing data in big datasets. These ensemble algorithms are a group of techniques that their decisions are combined in a way to optimize the execution of a specific algorithm [112]. Developing an ensemble involves of certain steps which are creating varied models and merging their estimates(see 0.0.2 Ensemble Generation). It is to be noted that ensemble techniques are best suited mostly where the highest possible accuracy is desired [113]. Before an ensemble is created there need to be a strategy in-order to build an ensemble that is as diverse as possible. This is because building the best ensemble method depends much on the problem that is being handled [114]. They are several ensemble strategies that are used, and these include but are not limited to Bagging, Boosting and Stacking.

Ensemble Generation The general ensemble algorithm creation which was formalized by [115] consists of two steps as stated above. The steps involve selecting points(creating varied models) and fitting coefficients(merging their estimates) which are explained in detail below.

1 selecting points $\{q_m\}_I^M$

1: $T_0(x) = 0$
2: For $m = 1 to M$
3: $q_m = \underset{q}{argmin} \sum_{i\epsilon_{S_m(\eta)}} L(y_i, T_{m-1}(X_i) + F(x_i; q))$
4: $F_m(x) = F(x; q_m)$
5: $T_m(x) = T_{m-1}(x) + \nu \cdot F_m(x)$
6: write$\{F_m(x)\}_I^M$

2 Choose Coefficients $\{c_m\}_O^M$

After all the base learners $\{F_m(x)\}_I^M = \{F(x; q_m)\}_I^M$ have been selected the coefficients are obtained by linear regression:
$\{\hat{c_m}\} = argmin \sum_{I=1}^{N} L\left(y_i, co + \sum_{m=1}^{M} cmF_m(x_i)\right) \lambda \cdot Q(c)$, where Q_c is the complexity penalty and λ represents the meta-parameter. The other three parameters L, η ν, L represent the loss function, η is responsible for data distribution and $S(\eta)$ represents a random sample that is the same size or less than the original data. If the values of η are smaller the diversity of the ensemble will increase, also, η has an effect on computing time. ν, regulates the alarms to the loss function.

The algorithm explains the start of an ensemble T_0 with a function(Line 1) which can be zero or any other constant. Then a leaner F_m is included into the process. $T_m - 1$ displays the ensemble of the base learners till $m - 1$. $q_m = argmin_q$... finds the lowest error base leaner on a selected data set. That is a base learner is chosen that when combining with other selected learners best approximates the response. The new base leaner is then added to the ensemble which is represented by F_m. After M base learner have been created the algorithm ends the process.

Bagging: This is a combination method where each ensemble is trained using dissimilar training sets which are generated by sampling the original set, choosing N items uniformly at random with replacement [116]. The missing values predictions of the algorithms are then combined by averaging or voting. One major high notes of bagging it is that it is a standout and simple ensemble methods to implement and has great execution.

AdaBoost: Boosting is the procedure of taking weak learning missing handling algorithms and turning them into strong learning algorithms. Similar to bagging, boosting also re-samples data to create ensembles, which are then combined by majority voting. However, similarities end there. Different variations of Boosting have been done and proved to be good as far as expectation exactness in an assortment of uses. Its major drawback is the slow training speed of boosted trees [117]. However, the highlight of AdaBoost is that it can be utilised to enhance the performances of other data mining algorithms regardless of their nature [118].

Stacking: Stacking is a mechanism that combines different types of models that have been learned in the task into one. The missing value predictions of different models gives an input on a meta-level classifier and the output of this meta classifier will be the final prediction [119]. The major component in stacking is the optimal features and the algorithm for learning at the meta-level [120]. It has been shown that with stacking the ensemble performs similar to choosing the best classifier from the ensemble by cross-validation. Instead of choosing one generalisation out of multiple generalisations, stacking combines them by using their output of base classifiers as inputs into the new space. Stacking then makes predictions from new space, stacking is considered to be an ensemble for further research in the context of base-level classifiers created by different learning algorithms [121].

Approaches in literature on missing values handling using ensemble methods are discussed in the following. Authors in[122], proposed a bootstrapping ensemble to model uncertainty and variety in the data that has high missingness. They performed an extensive evaluation of their approach by varying the missingness ratio of the missing data. Their results illustrated that bootstrapping is the most robust method for missing value imputation even at a high missingness ratio of up to 30 percent. However, for a small missingness ratio of 10 percent the ensemble strategy performed equivalently to other approaches but better than single imputation. Furthermore, the study was carried out using the MCAR missingness mechanism only, making their findings to be valid solely for this type of missingness.

Also, in another study [123] the authors proposed a Multiple Imputation Ensembles approach for handling with missing data in classification problems. They combined multiple imputation and ensemble techniques and compared two types of ensembles namely, bagging and stacking. The approach was termed robust as 20 benchmark datasets from the UCI machine learning repository were used. An increasing amount of missing data completely at random was simulated for all the data sets. It was reported that the approach performed well. However, it was not possible for the experiments results to be directly compared to other works on related work since different datasets and experimental set-ups were used.

Moreover, in [124], a new approach for missing data using a three-way ensemble algorithm based on the imputation result was proposed. The objects with no missing values were firstly clustered by using a clustering method, then missing objects were filled using mean attribute's of each cluster. The experimental results of the study on UCI machine learning repository data sets verify that the proposed algorithm was effective. However, like many other approaches in literature a missing value mechanism was not considered.

Also, in [125], the researchers developed a novel ensemble imputation approach named the missXGBoost imputation technique. The technique has proven to be suitable for

continuous attributes of medical applications. The missXGBoost method imputed plausible missing values in the original dataset and evaluated the classifier accuracy. The study experimental results demonstrated that the proposed imputation approach accuracy was better than the other traditional imputation methods. Furthermore, the method could be applied to high-dimensional mixed-type attributes data sets.

In another research a bagging and boosting ensemble algorithms as methods for handling missing data was proposed [126]. The proposed technique was compared with the existing methods by simulation and then applied to analyse a real large dataset to obtain realistic results. The researchers concluded that there is a lot of work to further experiment with their approach.

The following table below 2, presents a summary of different techniques in literature that used machine learning techniques to handle missing values. We present the general objective of the studies, the type of data set they used for their experiments, the missing mechanism followed and the limitations of the studies.

Table 2: A summary of various missing data techniques in machine learning

Ref.	DataSet	Performance Objective	Mechanism	Summary	Limitations
[127]	Balance, Breast, Glass, Bupa, Cmc, Iris, Housing, Ionosphere, wine	To study the influence of noise on missing value handling methods when noise and missing values distributed throughout the dataset	MCAR, MAR, MNAR	The technique proved that noise had a negative effect on imputation methods, particularly when the noise level is high.	Division of qualitative values may have been a problem
[88]	German, Glass(g2), heart-statlo, ionosphere, kr-vs-kp, labor, Pima-indians, sonar, balance-scale, iris, waveform, lymphography, vehicle, anneal, glass, satimage, image, zoo, LED, vowel, letter	Experimenting methods for handling incomplete training and test data for different missing data with various proportions and mechanisms.	MCAR, MAR	In this technique an understanding of the relative strengths and weaknesses of decision trees for missing value imputation was discussed.	The approach did not consider correlations between features.
[128]	Los Angeles ozone pollution and Simulated data	To study classification and regression problems using a variety of missing data mechanisms in order to compare the approaches on high dimensional problems.	MCAR, MAR	Here the authors tested the potential of imputation technique's dependence on the correlation structure of the data.	Random choice of missing values may have weakened the experiment consistency
[129]	Liver, Diabetis, Breast Cancer, Heart, WDSC, Sonar	Experimented on missing data handling using Random Forests and specifically analysed the impact of correlation of features on the imputation results.	MCAR, MAR, MNAR	The imputation approach was reported to be generally robust with performance improving when increasing correlation.	Random choice of missing values in MNAR could have weakened the consistency of the experiment
[130]	Wine , Simulated	To create an improved imputation algorithm for handling missing values.	MCAR, MAR, MNAR	Demonstrated the superiority of a new algorithm to existing imputation methods on accuracy of imputing missing data.	Features may have had different percentages of missing data, also MAR and MNAR may have been weakened.

Table 2 – continued from previous page

Ref.	DataSet	Performance Objective	Mechanism	Summary	Limitations
[131]	Iris, Wine, Voting, Tic-Tiac-Toe, Hepatitis	To propose a novel technique to impute missing values based on feature relevance	MCAR, MAR	The approach employed mutual information to measure feature relevance and proved to reduce classification bias.	Random choice of missing values may have weakened the experiment consistency
[132]	Pima Indian Diabetes dataset	To experiment on missing values approach that takes into account feature relevance		The results of the technique proved that the hybrid algorithm was better than the existing methods in terms of accuracy, RMSE and MAE .	Missing values mechanism was not considered.
[17]	Iris , Voting, Hepatitis	Proposed an iterative KNN that took into account the presence of the class labels	MCAR, MAR	The technique considered class labels and proved to perform good against other imputation methods.	The approach has not been theoretically proven to converge, though it was empirically shown
[77]	Camel, Ant, Ivy, Arc, Pcs, Mwl, KC3, Mc2	To develop a novel incomplete-instance based imputation approach that utilized cross-validation to improve the parameters for each missing value.	MCAR, MAR	The study demonstrated that their approach was superior to other missing values approaches.	
[133]	Blood, breast-cancer, ecoli, glass, ionosphere, iris, Magic, optdigits, pendigits, pima, segment, sonar, waveform, wine, yeast, balance-scale, Car, chess-c, chess-m, CNAE-9, lymphography, mushroom, nursery, promoters, SPECT, tic-tac-toe, abalone, acute, card, contraceptive, German, heart, liver, zoo	To develop a missing handling approach is introduced with effective imputation results.	MCAR	The method was based on calculating the class center of every class and using the distances between it and the observed data to define a threshold for imputation.The method performed better and had less imputation time.	Only one missing mechanism was implemented
[134]	Groundwater	Developed a multiple imputation method that can handle the missingness in ground water dataset with high rate of missing values .	MAR	Here the technique used to handle the missing values, was chosen looking at its ability to consider the relationships between the variables of interest.	There was no prior knowledge on the label of missing data which may have provided difficulty when performing imputation
[135]	Dukes' B colon cancer, the Mice Protein Expression and Yeast	Developed a novel hybrid Fuzzy C means Rough parameter missing value imputation method.		The technique handled the vagueness and coarseness in the dataset and proved to produce better imputation results.	There was no report of missing values mechanisms used for the experiment.
[136]	Forest fire,Glass, Housing, Iris, MPG, MV, Stocks, Wine	The method proposed a variant of the forward stage-wise regression algorithm for data imputation by modelling the missing values as random variables following a Gaussian mixture distribution. Categorical		The method proved to be effective compared to other approaches that combined standard missing data approaches and the original FSR algorithm.	There was no report of missing values mechanisms used for the experiment.

Table 2 – continued from previous page

Ref.	DataSet	Performance Objective	Mechanism	Summary	Limitations
[137]	Weather dataset	This method applied four(Likewise, Multiple imputation,KNN,MICE) missing data handling methods to the training data before classification		Of the imputation methods applied the authors concluded that the most effective missing data imputation method for photovoltaic forecasting was the KNN method.	There was no report of missing values mechanisms used for the experiment
[138]	Air quality data	To make time series prediction for missing values using three machine learning algorithms and identify the best method.		The study concluded that deep learning performed better when data was large and machine learning models produced better results when the data was less.	Heavy costs in time consumption and computational powers for training when implementing their most effective method(deep learning).
[139]	Traumatic Brain Injury and Diabetes	To demonstrate how performance varies with different missing value mechanisms and the imputation method used and further demonstrate how MNAR is an important tool to give confidence that valid results are obtained using multiple imputation and complete case analysis.	MCAR, MAR, MNAR	The study showed that both complete case analysis and multiple imputation can produce unbiased results under more conditions.	The method was limited by the absence of nonlinear terms in the substantive models.
[140]	Grades Dataset	To develop a new decision tree approach for missing data handling	MCAR, MAR, MNAR	The method produced a higher accuracy compared to other missing values handling techniques and had more interpretable classifier.	The algorithm suffered from a weakness when the gating variable had no predictive power.
[141]	Air Pressure System data	The study proposed a sorted missing percentages approach for filtering attributes when building machine learning classification model using sensor readings with missing data.		The technique proved to be effective for scenarios dealing with missing data in industrial sensor data analysis.	The proposed approach could not meet the needs of automation.
[141]	Abalone and Boston Housing	To experiment the reliability of missing value handling at not missing at random.	MAR	The results of the study indicated that the approach achieved satisfactory performance in solving the lower incomplete problem compared to other six methods.	The approach did not consider any missingness rate which may have affected the analysis.

Performance metrics for missing data imputation

The performance evaluation of different missing values approaches in machine learning problems can be done using different criteria, on this section we discuss the most commonly used which are, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Area under the curve (AUC).

Mean Absolute Error(MAE)

MAE measures the average difference between imputed values and true values defined as:

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y_i}| \tag{18}$$

Mean Squared Error(MSE)

While MSE is equal to the sum of variance and squared predicted missing value as in the following equation:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y_i})^2 \tag{19}$$

Root Mean Square Error(RMSE)

RMSE computes the difference in imputed values and actual values as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i-1}^{m}(y_i - \bar{y_i})^2} \tag{20}$$

MSE measures the average squared difference between the predicted missing values and the actual value, while RMSE represents the standard deviation of the differences between estimated missing values and observed values. Where m is the number of observations, y_i is the observed values and $\bar{y}_i$ is the estimated missing value. A small value as an output for these performance metrics means that the estimated value is close to the real value.

Area Under The Curve(AUC)

AUC is the representation of the degree or measure of separability and is used as a summary of the Root Receiver Operator Characteristic (ROC) curve, which is curve is a visualisation graph representing imputation performance [142]. The AUC is represented by the true positive rate (TPR) and the false positive rate (FPR). Where the TPR is the proportion of correctly imputed positives of all positives and the TPR is the proportion of all negatives that are wrongly imputed as positives [143]. The true positive rate and the false positive rate are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FPR = \frac{FP}{FP + TN} \tag{22}$$

The major advantages of the MSE and RMSE is that they provide a quadratic loss function. Also, uncertainty in forecasting is measured when they are used. However, MSE and RMSE are highly influenced by extreme values [144]. While, MAE is not influenced by extreme values, also a more natural measure and unambiguous [145]. Most studies in research are found to mostly use the RMSE for missing value imputation evaluation [146–148]. Although some studies have proposed valid evidence against the use of RMSE in favor of MAE due to its less sensitive to extreme values [149]. The authors further advised against the reporting of RMSE in literature and strongly recommended the use of MAE [145, 149]. However, [144] partially disputed the conclusions and introduced arguments against avoiding RMSE. They contended that RMSE was appropriate to represent model performance than the MAE. The AUC like other performance measures also has its advantages, it allows for a visualised graphical representation of imputation performance and is also unaffected by abnormal distributions in the population and decision criterion [150]. However, actual decision thresholds are usually not represented by AUC graph and it overlooks the probability of predicted values and the goodness-of-fit of the model [151]. Discussions on which metric to use in literature have proven that performance measures are not equivalent to each other, and one cannot easily derive the value of one from another. Nonetheless, all distance measurements (MSE, RMSE, MAE and AUC) help to quantify the accuracy of the estimated solution compared to the actual non-missing data and an appropriate method must be selected for the most appropriate analysis for the question being addressed.

Comparisons

In this section, we discuss observations made, and present a comparative analysis on performance matrices, publications made and the year of publication for different researches.

Evaluation Metrics

Table 3 shows details of different selected articles that were researched on missing data handling using different techniques and the journals, books, conference they were published on. We selected articles in Table 4 for metrics used to evaluate different missing values handling approaches. The selection is based on whether the article covers the most popular evaluation methods.

Table 3: Details of selected articles for missing values handling.

Citation	Year	Publisher	Article	Journal/Conference/Book Chapter
[152]	2020	Applied Science	Missing Value Imputation in Stature Estimation by Learning Algorithms Using Anthropometric Data: A Comparative Study	Multidisciplinary Digital Publishing Institute
[141]	2020	Applied Science	Evaluating Machine Learning Classification Using Sorted Missing Percentage Technique Based on Missing Data	Multidisciplinary Digital Publishing Institute
[153]	2020	Biometrical Journal	Multiple imputation methods for handling missing values in longitudinal studies with sampling weights: Comparison of methods implemented in Stata	Wiley Online Library
[154]	2019	Applied Artificial Intelligence	Comparison of performance of data imputation methods for numeric dataset	Taylor and Francis
[9]	2006	Elsevier	A gentle introduction to imputation of missing values	Journal of clinical epidemiology
[129]	2017	Elsevier	Adjusted weight voting algorithm for random forests in handling missing values	Pattern Recognition
[63]	2017	Elsevier	kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data	Knowledge-Based Systems
[155]	2021	Elsevier	Ground PM2. 5 prediction using imputed MAIAC AOD with uncertainty quantification	Environmental Pollution
[156]	2021	Elsevier	A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system	Expert Systems with Applications
[157]	2021	Elsevier	Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018)	Multidisciplinary Digital Publishing Institute
[158]	2021	Elsevier	HA new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for internet of vehicle	Neurocomputing
[114]	2006	IEEE	Ensemble based systems in decision making	IEEE Circuits and systems magazine
[159]	2010	IEEE	Missing Value Estimation for Mixed-Attribute Data Sets	IEEE Transactions on Knowledge and Data Engineering
[160]	2014	IEEE	Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge	IEEE Signal Processing Magazine
[2]	2014	IEEE	Handling missing data problems with sampling methods	2014 International Conference on Advanced Networking Distributed Systems and Applications

Table 3 – continued from previous page

Citation	Year	Publisher	Article	Journal/Conference/Book Chapter
[126]	2018	IEEE	An imputation method for missing data based on an extreme learning machine auto-encoder	IEEE ACCESS
[161]	2018	IEEE	A data imputation model in phasor measurement units based on bagged averaging of multiple linear regression	IEEE ACCESS
[162]	2018	IEEE	Missing network data a comparison of different imputation methods	2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)
[163]	2018	IEEE	MIAEC: Missing data imputation based on the evidence chain	IEEE ACCESS
[164]	2018	IEEE	A survey on data imputation techniques: Water distribution system as a use case	IEEE ACCESS
[165]	2019	IEEE	Missing Values Estimation on Multivariate Dataset: Comparison of Three Type Methods Approach	International Conference on Information and Communications Technology (ICOIACT)
[125]	2019	IEEE	A Novel Algorithm for Missing Data Imputation on Machine Learning	2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)
[166]	2020	IEEE	Approaches to Dealing With Missing Data in Railway Asset Management	IEEE ACCESS
[167]	2020	IEEE	Traffic Data Imputation and Prediction: An Efficient Realization of Deep Learning	IEEE ACCESS
[168]	2020	IEEE	Iterative Robust Semi-Supervised Missing Data Imputation	IEEE ACCESS
[169]	2021	IEEE	Missing network data a comparison of different imputation methods Neighborhood-aware autoencoder for missing value imputation	2020 28th European Signal Processing Conference (EUSIPCO)
[170]	2021	IEEE	Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set	IEEE Transactions on Fuzzy Systems
[59]	2016	SAGE Publications	Multiple imputation in the presence of high-dimensional data	Statistical Methods in Medical Research
[171]	2020	Sensors	A Method for Sensor-Based Activity Recognition in Missing Data Scenario	Multidisciplinary Digital Publishing Institute
[35]	2012	Springer	Analysis of missing data	Missing data
[68]	2015	Springer	CKNNI: an improved knn-based missing value handling technique	International Conference on Intelligent Computing
[131]	2015	Springer	Missing data imputation by K nearest neighbours based on grey relational structure and mutual information	Applied Intelligence
[66]	2016	Springer	Nearest neighbor imputation algorithms: a critical evaluation	BMC medical informatics and decision making
[108]	2017	Springer	Multiple imputation and ensemble learning for classification with incomplete data	Intelligent and Evolutionary Systems
[71]	2018	Springer	NS-kNN: A modified k-nearest neighbors approach for imputing metabolomics data	Metabolomics
[138]	2019	Springer	Analysis of interpolation algorithms for the missing values in IoT time series: a case of air quality in Taiwan	The Journal of Super computing
[42]	2020	Springer Open	SICE: an improved missing data imputation technique	Journal of Big Data
[140]	2020	Springer	BEST: a decision tree algorithm that handles missing values	Computational Statistics

Table 3 – continued from previous page

Citation	Year	Publisher	Article	Journal/Conference/Book Chapter
[172]	2020	Springer	A new multi-view learning machine with incomplete data	Pattern Analysis and Applications
[173]	2021	Springer	Multistage Model for Accurate Prediction of Missing Values Using Imputation Methods in Heart Disease Dataset	Innovative Data Communication Technologies and Application
[18]	2021	Springer	A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data	Soft Computing
[174]	2021	Springer	An Exploration of Online Missing Value Imputation in Non-stationary Data Stream	SN Computer Science
[175]	2021	Springer	Data Imputation in Wireless Sensor Network Using Deep Learning Techniques	Data Analytics and Management
[176]	2020	Sustainable and Resilient Infrastructure	Handling incomplete and missing data in water network database using imputation methods	Taylor and Francis

Table 4: Qualitative comparison between different missing data techniques in machine learning based on the performance metrics adopted.

Publication	Performance Metrics			
	RMSE	MAE	MSE	AUC
[128]	×	×	✓	×
[177]	×	×	✓	✓
[77]	✓	×	×	×
[129]	×	×	×	✓
[133]	✓	✓	×	×
[135]	✓	×	×	×
[137]	✓	×	×	×
[138]	×	✓	✓	×
[131]	✓	×	×	×
[132]	✓	✓	×	×
[130]	✓	×	×	×
[48]	×	×	✓	×
[141]	×	×	×	✓
[140]	×	×	×	✓
[173]	✓	×	×	×
[174]	✓	×	×	×
[155]	✓	×	×	×
[157]	✓	✓	×	×
[169]	✓	×	×	×
[19]	✓	×	×	×
[178]	×	×	✓	✓
[41]	×	×	✓	✓

Experimental evaluation on Machine Learning Methods

An experimental evaluation on missing values techniques is beyond the scope of this work. However, we selected some of the most representative machine learning techniques to show experimental results on two datasets. Considering the possible variability on performances of algorithms, the experiment was done on more than one algorithm based on the Iris and ID fan datasets. The Iris dataset is a very popular dataset which was originally published at UCI Machine Learning Repository introduced by [179], for an application of discriminant analysis for three species of Iris flowers (setosa, versicolor, and virginica), having four variables being length and width of the sepal and the petal (in cm). We also experimented on an Induced draft fan (ID fan) dataset from a local coal-fired power plant where real data of a coal power plant fan system was recorded. The dataset contains readings for the month of February 2021 of a single unit of the power plant. The ID fan vibrations are measured by sensors and were recorded by the technicians every 4 hours when the plant was running.

These variables specifically consist of bearing vibrations and temperatures, at the fan non-drive end (FNDE) and fan drive end (FDE), motor temperatures and vibrations, at the motor non-drive end (MNDE) and motor drive end (MDE).The values of the ID fan are recorded as part of the daily power plant monitoring system. Both the Iris and Id Fan datasets contain 150 instances with no missing values.Our method simulates the missing values on sepal length and petal width of the Iris data and the Vibrations on the ID fan data at a ratio of 5%, 10% and 15%. RMSE performance measure was then used to help quantify the accuracy of the estimated values compared to the actual non-missing data.

After simulation of missing values, KNN imputation was implemented to replace the missing values. Firstly when implementing the imputation method, the nearest neighbors (K) must be chosen. The value of K was chosen based on experimental results starting with K=1 and stopped at K=5, the best accurate estimation value of K was then used for the experiment which was K=4.

The Random Forests algorithm was then implemented for performance comparison with the KNN.

Figure 1: Comparison of KNN and RBF Imputed values with the actual values on Sepal Length at 15% ratio

Figure 2: Comparison of KNN and RBF Imputed values with the actual values on Petal width at 15%

Table 5: RMSE of KNN and RBF imputation at different ratios on the Iris dataset.

Missing Ratio%	**KNN**	**RBF**
5	0.6693	0.6486
10	0.2382	0.2860
15	0.1932	0.2578

Figure 3: Comparison of KNN and RBF Imputed and actual values at 15% on the Id fan dataset

Table 6: RMSE of KNN and RBF imputation at different ratios on the ID fan dataset.

Missing Ratio%	**KNN**	**RBF**
5	0.2099	0.0549
10	0.1581	0.0416
15	0.1487	0.0654

Table 5 and 6 represents the RMSE of the KNN and RBF algorithms at different imputation ratios on the Iris and ID fan datasets. The experiment demonstrated that the KNN imputation generally performed better than the RBF imputation on the Iris dataset. The RBF showed to only perform better than the KNN when there was a small ratio of missing values at 5% on the Iris dataset and the performance gradually decreased when the missing percentage increased. However, the RBF performed better than the KNN on the ID fan data sets at all missing value percentages.

Conclusion and Future Work

Most of the real-world data analysis-based research face the problem of data containing missing values. This paper discusses the problem of missing values, including missing data mechanisms (MCAR, MAR, and MNAR), missingness types and a considerable number of missing data handling approaches, for different applications and scenarios. We also illustrate missing data approaches in

specific contexts and how they work. Furthermore, an in-depth analysis of evaluation techniques was done and it is notable that literature on handling missing values mostly uses RSME for evaluation. There are also a number of studies that evaluate by using imputation for data pre-processing during classification and compare the algorithm accuracy before and after imputation. We also experimented on the KNN and RBF algorithms for imputation on the Iris and ID fan datasets. KNN imputation performed better than the RBF imputation using RMSE as an evaluation measure on the Iris data and the RBF performed better than the KNN on the ID fan data. This has lead to a conclusion that, the precision and accuracy of machine learning imputation algorithms depend strongly on the type of data being analysed, and that there is no clear indication that favours one method over the other. There is also limited research on missing data imputation on big data sets and high dimensional datasets. Therefore, further work is needed to explore the possibilities of new methods of handling missing data in big data using optimized approaches.

Appendix

The tables explains some notations used on the manuscript. Table 7 illustrates the summary of the notations and definitions used on the paper.

Table 7: Summary of notation and definitions.

Notation	Definition
b	The Bias
$Dist_{xy}$	The euclidean distance
$f(y_{obs})$	The complete data in the data set
H	The separating hyper-plane
k	The data attributes
m	The number of observations
n	The number of observed data
p	The probability of missing data
q	The vector indicating the missingness association
R	The missing value matrix
v_j	Attributes containing missing data
w	The weight vector
w_j	Nearest neighbours
x	The input vector
x_i.	The error terms for un-predicted determinants of $\bar{y}$
X_k	Mean estimation
Y	The matrix of an entire data set
Y_m	The missing Data in R
Y_o	The observed data in R
$\bar{y}$	The predicted data

Abbreviations

AUC: Area under the curve, CART: Classification and Regression Trees, FNDE:Fan Non-Drive End, FDE:Fan Drive End, MEA: Mean Absolute Error, MDE: Motor Drive End, MSE: Mean Squared Error MNDE:Motor Non-Drive End, KNN: K nearest neighbor,MAR:, Missing at Random, MCAR: Missing Completely at Random, MNAR:Missing Not at Random, RBF: Random Forests, RMSE: Root Mean Squared Error, UCI:University of California, SVM: Support Vector Machines.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The availability one data sources is explained in the manuscript with a reference.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

TE designed the study and developed the experiments, and led the writing of the paper; TM provided the concept, consultation and reviewed the paper; DM reviewed the paper; TS developed the tables and found papers used in the survey. MB and OT reviewed and edited the paper. All authors read and approved the final manuscript.

Funding
This work received a grant from the Botswana International University of Science and Technology.

Acknowledgements
Not applicable.

Additional Files
Figures
The additional figures which are to be part of the final manuscript are included in a separate .png documents.

References

1. Suthar, B., Patel, H., Goswami, A.: A survey: classification of imputation methods in data mining. International Journal of Emerging Technology and Advanced Engineering **2**(1), 309–12 (2012)
2. Houari, R., Bounceur, A., Tari, A.K., Kecha, M.T.: Handling missing data problems with sampling methods. In: 2014 International Conference on Advanced Networking Distributed Systems and Applications, pp. 99–104 (2014). IEEE
3. McKnight, P.E., McKnight, K.M., Sidani, S., Figueredo, A.J.: Missing Data: A Gentle Introduction. Guilford Press, ??? (2007)
4. Ayilara, O.F., Zhang, L., Sajobi, T.T., Sawatzky, R., Bohm, E., Lix, L.M.: Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and quality of life outcomes **17**(1), 106 (2019)
5. Kang, H.: The prevention and handling of the missing data. Korean journal of anesthesiology **64**(5), 402 (2013)
6. Ludbrook, J.: Outlying observations and missing values: how should they be handled? Clinical and experimental pharmacology & physiology **35**(5-6), 670–678 (2008)
7. Zhang, Z.: Missing values in big data research: some basic skills. Annals of translational medicine **3**(21) (2015)
8. Langkamp, D.L., Lehman, A., Lemeshow, S.: Techniques for handling missing data in secondary analyses of large surveys. Academic pediatrics **10**(3), 205–210 (2010)
9. Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G.: A gentle introduction to imputation of missing values. Journal of clinical epidemiology **59**(10), 1087–1091 (2006)
10. Shawe-Taylor, J., Cristianini, N., *et al.*: Kernel Methods for Pattern Analysis. Cambridge university press, ??? (2004)
11. Graham, J.W.: Missing data analysis: Making it work in the real world. Annual review of psychology **60**, 549–576 (2009)
12. Baraldi, A.N., Enders, C.K.: An introduction to modern missing data analyses. Journal of school psychology **48**(1), 5–37 (2010)
13. Muhammad, I., Yan, Z.: Supervised machine learning approaches: A survey. ICTACT Journal on Soft Computing **5**(3) (2015)
14. Nithya, C., Saravanan, V.: A study of machine learning techniques in data mining. International Scientific Refereed Research Journal (2018)
15. Aydilek, I.B., Arslan, A.: A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences **233**, 25–35 (2013)
16. Lin, J., Li, N., Alam, M.A., Ma, Y.: Data-driven missing data imputation in cluster monitoring system based on deep neural network. Applied Intelligence **50**(3), 860–877 (2020)
17. Choudhury, A., Kosorok, M.R.: Missing data imputation for classification problems. arXiv preprint arXiv:2002.10709 (2020)
18. Al-Helali, B., Chen, Q., Xue, B., Zhang, M.: A new imputation method based on genetic programming and weighted knn for symbolic regression with incomplete data. Soft Computing, 1–20 (2021)
19. Peng, D., Zou, M., Liu, C., Lu, J.: Resi: A region-splitting imputation method for different types of missing data. Expert Systems with Applications **168**, 114425 (2021)
20. Fan, J., Han, F., Liu, H.: Challenges of big data analysis. National science review **1**(2), 293–314 (2014)
21. Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S.: A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing **2016**(1), 1–16 (2016)
22. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data vol. 793. John Wiley & Sons, ??? (2019)
23. De Leeuw, E.D., Hox, J.J., Huisman, M.: Prevention and treatment of item nonresponse. Journal of Official Statistics **19**, 153–176 (2003)
24. Berglund, P., Heeringa, S.G.: Multiple Imputation of Missing Data Using SAS. SAS Institute, ??? (2014)
25. Demirtas, H.: Flexible imputation of missing data. Journal of Statistical Software **85**(1), 1–5 (2018)
26. Lacerda, M., Ardington, C., Leibbrandt, M.: Sequential regression multiple imputation for incomplete multivariate data using markov chain monte carlo (2007)
27. Liu, C.: Missing data imputation using the multivariate t distribution. Journal of multivariate analysis **53**(1), 139–158 (1995)
28. Dong, Y., Peng, C.-Y.J.: Principled missing data methods for researchers. SpringerPlus **2**(1), 222 (2013)
29. Chen, Y.-C.: Pattern graphs: a graphical approach to nonmonotone missing data. arXiv preprint arXiv:2004.00744 (2020)
30. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
31. Gómez-Carracedo, M., Andrade, J., López-Mahía, P., Muniategui, S., Prada, D.: A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems **134**, 23–33 (2014)
32. Yang, X., Li, J., Shoptaw, S.: Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. Statistics in medicine **27**(15), 2826–2849 (2008)
33. Grittner, U., Gmel, G., Ripatti, S., Bloomfield, K., Wicki, M.: Missing value imputation in longitudinal measures of alcohol consumption. International journal of methods in psychiatric research **20**(1), 50–61 (2011)

34. Dantan, E., Proust-Lima, C., Letenneur, L., Jacqmin-Gadda, H.: Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts. The International Journal of Biostatistics **4**(1) (2008)
35. Graham, J.W.: Analysis of missing data. In: Missing Data, pp. 47–69. Springer, ??? (2012)
36. Soley-Bori, M.: Dealing with missing data: Key assumptions and methods for applied analysis. Boston University **23** (2013)
37. Williams, R.: Missing data Part 1: Overview, traditional methods. University of Notre Dame (2015)
38. Allison, P.D.: Missing Data vol. 136. Sage publications, ??? (2001)
39. Kim, J.-O., Curry, J.: The treatment of missing data in multivariate analysis. Sociological Methods & Research **6**(2), 215–240 (1977)
40. García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., Verleysen, M.: K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing **72**(7-9), 1483–1493 (2009)
41. Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine **50**(2), 105–115 (2010)
42. Khan, S.I., Hoque, A.S.M.L.: Sice: an improved missing data imputation technique. Journal of Big Data **7**(1), 1–21 (2020)
43. Song, Q., Shepperd, M.: Missing data imputation techniques. International journal of business intelligence and data mining **2**(3), 261–291 (2007)
44. Yu, L., Liu, L., Peace, K.E.: Regression multiple imputation for missing data analysis. Statistical Methods in Medical Research, 0962280220908613 (2020)
45. Alexopoulos, E.C.: Introduction to multivariate regression analysis. Hippokratia **14**(Suppl 1), 23 (2010)
46. De Waal, T., Pannekoek, J., Scholtus, S.: Handbook of Statistical Data Editing and Imputation vol. 563. John Wiley & Sons, ??? (2011)
47. Sherwood, B., Wang, L., Zhou, X.-H.: Weighted quantile regression for analyzing health care cost data with missing covariates. Statistics in medicine **32**(28), 4967–4979 (2013)
48. Crambes, C., Henchiri, Y.: Regression imputation in the functional linear model with missing values in the response. Journal of Statistical Planning and Inference **201**, 103–119 (2019)
49. Siswantining, T., Soemartojo, S.M., Sarwinda, D., *et al.*: Application of sequential regression multivariate imputation method on multivariate normal missing data. In: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), pp. 1–6 (2019). IEEE
50. Andridge, R.R., Little, R.J.: A review of hot deck imputation for survey non-response. International statistical review **78**(1), 40–64 (2010)
51. Cheema, J.R.: A review of missing data handling methods in education research. Review of Educational Research **84**(4), 487–508 (2014)
52. Sullivan, D., Andridge, R.: A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. Computational Statistics & Data Analysis **82**, 173–185 (2015)
53. Christopher, S.Z., Siswantining, T., Sarwinda, D., Bustaman, A.: Missing value analysis of numerical data using fractional hot deck imputation. In: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), pp. 1–6 (2019). IEEE
54. Lin, W.-C., Tsai, C.-F.: Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review **53**(2), 1487–1509 (2020)
55. Rubin, L.H., Witkiewitz, K., Andre, J.S., Reilly, S.: Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. Journal of Undergraduate Neuroscience Education **5**(2), 71 (2007)
56. Delalleau, O., Courville, A., Bengio, Y.: Efficient em training of gaussian mixtures with missing data. arXiv preprint arXiv:1209.0521 (2012)
57. Uusitalo, L., Lehikoinen, A., Helle, I., Myrberg, K.: An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software **63**, 24–31 (2015)
58. Nguyen, C.D., Carlin, J.B., Lee, K.J.: Model checking in multiple imputation: an overview and case study. Emerging themes in epidemiology **14**(1), 8 (2017)
59. Zhao, Y., Long, Q.: Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research **25**(5), 2021–2035 (2016)
60. Huque, M.H., Carlin, J.B., Simpson, J.A., Lee, K.J.: A comparison of multiple imputation methods for missing data in longitudinal studies. BMC medical research methodology **18**(1), 168 (2018)
61. Horton, N.J., Lipsitz, S.R., Parzen, M.: A potential for bias when rounding in multiple imputation. The American Statistician **57**(4), 229–232 (2003)
62. de Goeij, M.C., van Diepen, M., Jager, K.J., Tripepi, G., Zoccali, C., Dekker, F.W.: Multiple imputation: dealing with missing data. Nephrology Dialysis Transplantation **28**(10), 2415–2420 (2013)
63. Maillo, J., Ramírez, S., Triguero, I., Herrera, F.: knn-is: An iterative spark-based design of the k-nearest neighbors classifier for big data. Knowledge-Based Systems **117**, 3–15 (2017)
64. Amirteimoori, A., Kordrostami, S.: A euclidean distance-based measure of efficiency in data envelopment analysis. Optimization **59**(7), 985–996 (2010)
65. Gimpy, M.: Missing value imputation in multi attribute data set. International Journal of Computer Science and Information Technologies **5**(4), 1–7 (2014)
66. Beretta, L., Santaniello, A.: Nearest neighbor imputation algorithms: a critical evaluation. BMC medical informatics and decision making **16**(3), 74 (2016)
67. Acuna, E., Rodriguez, C.: The treatment of missing values and its effect on classifier accuracy. In: Classification, Clustering, and Data Mining Applications, pp. 639–647. Springer, ??? (2004)
68. Jiang, C., Yang, Z.: Cknni: an improved knn-based missing value handling technique. In: International Conference on Intelligent Computing, pp. 441–452 (2015). Springer

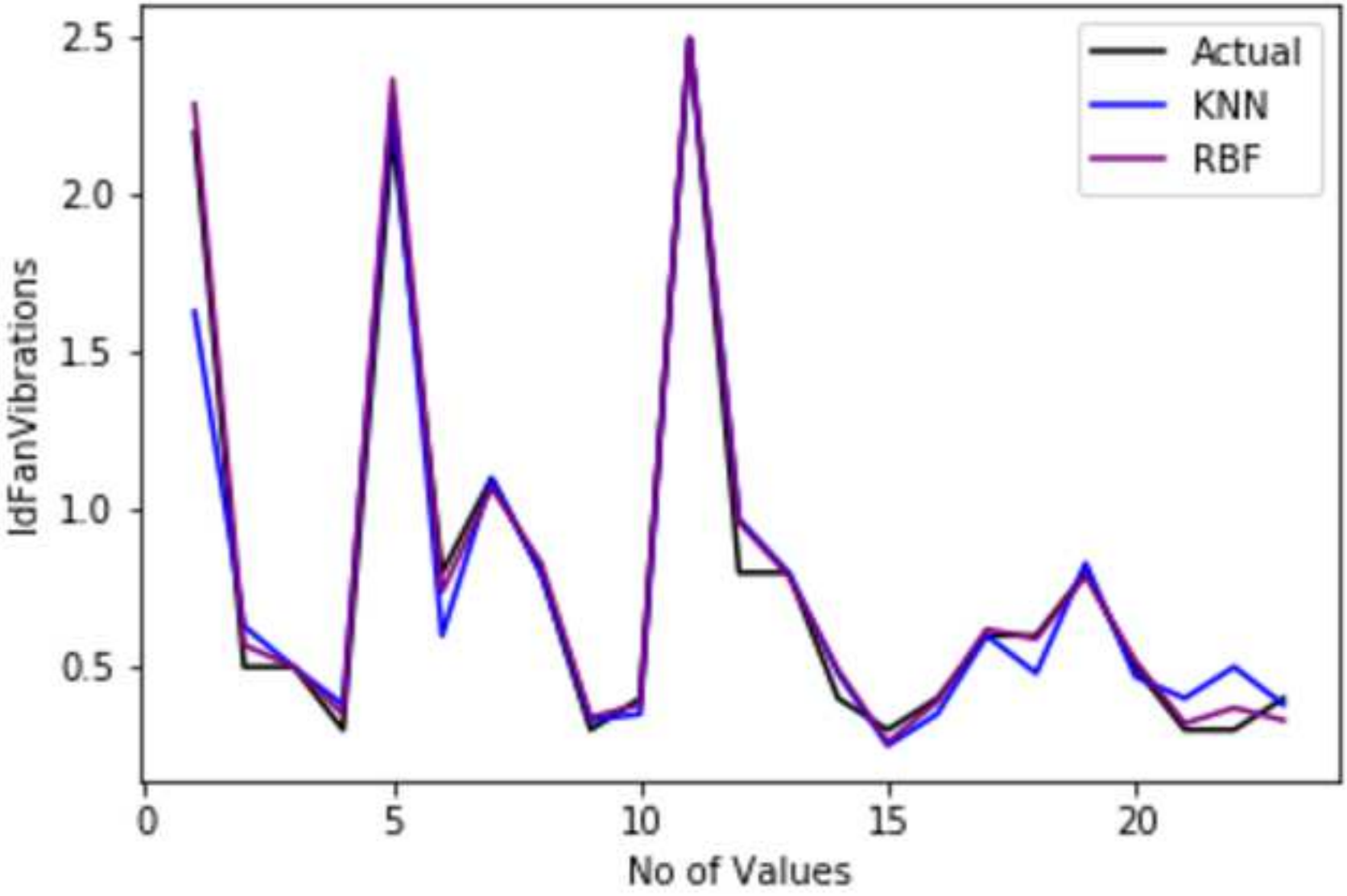
69. Sun, B., Ma, L., Cheng, W., Wen, W., Goswami, P., Bai, G.: An improved k-nearest neighbours method for traffic time series imputation. In: 2017 Chinese Automation Congress (CAC), pp. 7346–7351 (2017). IEEE
70. He, Y., Pi, D.-c.: Improving knn method based on reduced relational grade for microarray missing values imputation. IAENG International Journal of Computer Science **43**(3), 1–7 (2016)
71. Lee, J.Y., Styczynski, M.P.: Ns-knn: A modified k-nearest neighbors approach for imputing metabolomics data. Metabolomics **14**(12), 153 (2018)
72. Cheng, D., Zhang, S., Deng, Z., Zhu, Y., Zong, M.: knn algorithm with data-driven k value. In: International Conference on Advanced Data Mining and Applications, pp. 499–512 (2014). Springer
73. Meesad, P., Hengpraprohm, K.: Combination of knn-based feature selection and knnbased missing-value imputation of microarray data. In: 2008 3rd International Conference on Innovative Computing Information and Control, pp. 341–341 (2008). IEEE
74. Pujianto, U., Wibawa, A.P., Akbar, M.I., *et al.*: K-nearest neighbor (k-nn) based missing data imputation. In: 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 83–88 (2019). IEEE
75. Zhu, M., Cheng, X.: Iterative knn imputation based on gra for missing values in tplms. In: 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), vol. 1, pp. 94–99 (2015). IEEE
76. Huang, J., Sun, H.: Grey relational analysis based k nearest neighbor missing data imputation for software quality datasets. In: 2016 IEEE International Conference on Software Quality, Reliability and Security (QRS), pp. 86–91 (2016). IEEE
77. Huang, J., Keung, J.W., Sarro, F., Li, Y.-F., Yu, Y.-T., Chan, W., Sun, H.: Cross-validation based k nearest neighbor imputation for software quality datasets: An empirical study. Journal of Systems and Software **132**, 226–252 (2017)
78. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence **17**(5-6), 519–533 (2003)
79. De Silva, H., Perera, A.S.: Missing data imputation using evolutionary k-nearest neighbor algorithm for gene expression data. In: 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 141–146 (2016). IEEE
80. Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D.: Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST) **8**(3), 1–19 (2017)
81. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., Yumei, C.: A svm regression based approach to filling in missing values. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 581–587 (2005). Springer
82. Pelckmans, K., De Brabanter, J., Suykens, J.A., De Moor, B.: Handling missing values in support vector machine classifiers. Neural Networks **18**(5-6), 684–692 (2005)
83. Stewart, T.G., Zeng, D., Wu, M.C.: Constructing support vector machines with missing data. Wiley Interdisciplinary Reviews: Computational Statistics **10**(4), 1430 (2018)
84. Smola, A.J., Vishwanathan, S., Hofmann, T.: Kernel methods for missing variables. In: AISTATS (2005). Citeseer
85. Ghazanfar, M.A., Prugel, A.: The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations. Informatica **37**(1) (2013)
86. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning, pp. 137–142 (1998). Springer
87. Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D.: Max-margin classification of data with absent features. Journal of Machine Learning Research **9**(Jan), 1–21 (2008)
88. Twala, B.: An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence **23**(5), 373–405 (2009)
89. Rokach, L.: Decision forest: Twenty years of research. Information Fusion **27**, 111–125 (2016)
90. Rahman, M.G., Islam, M.Z.: Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. Knowledge-Based Systems **53**, 51–65 (2013)
91. Gimpy, D., Rajan Vohra, M.: Estimation of missing values using decision tree approach. Int J Comput Sci Inf Technol **5**(4), 5216–5220 (2014)
92. Rahman, G., Islam, Z.: A decision tree-based missing value imputation technique for data pre-processing. In: Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, pp. 41–50 (2011)
93. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC press, ??? (1984)
94. Phyu, T.N.: Survey of classification techniques in data mining. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1 (2009)
95. Gavankar, S., Sawarkar, S.: Decision tree: Review of techniques for missing values at training, testing and compatibility. In: 2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS), pp. 122–126 (2015). IEEE
96. Tang, F., Ishwaran, H.: Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal **10**(6), 363–377 (2017)
97. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
98. Stekhoven, D.J.: missforest: Nonparametric missing value imputation using random forest. Astrophysics Source Code Library, 1505 (2015)
99. Pantanowitz, A., Marwala, T.: Missing data imputation through the use of the random forest algorithm. In: Advances in Computational Intelligence, pp. 53–62. Springer, ??? (2009)
100. Hong, S., Lynn, H.S.: Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC medical research methodology **20**(1), 1–12 (2020)
101. Gajawada, S., Toshniwal, D.: Missing value imputation method based on clustering and nearest neighbours. International Journal of Future Computer and Communication **1**(2), 206–208 (2012)
102. Bhaduri, A., Bhaduri, A.: Color image segmentation using clonal selection-based shuffled frog leaping algorithm. In: 2009 International Conference on Advances in Recent Technologies in Communication and

Computing, pp. 517–520 (2009). IEEE
103. Zhang, S., Zhang, J., Zhu, X., Qin, Y., Zhang, C.: Missing value imputation based on data clustering. In: Transactions on Computational Science I, pp. 128–138. Springer, ??? (2008)
104. Besay Montesdeoca, J.L., Maillo, J., Garcıa-Gil, D., Garcıa, S., Herrera, F.: A first approach on big data missing values imputation (2019)
105. Zhang, Z., Fang, H., Wang, H.: Multiple imputation based clustering validation (miv) for big longitudinal trial data with missing values in ehealth. Journal of medical systems **40**(6), 146 (2016)
106. Zhang, C., Ma, Y.: Ensemble Machine Learning: Methods and Applications. Springer, ??? (2012)
107. Zhang, X.-F., Ou-Yang, L., Yang, S., Zhao, X.-M., Hu, X., Yan, H.: Enimpute: imputing dropout events in single-cell rna-sequencing data via ensemble learning. Bioinformatics **35**(22), 4827–4829 (2019)
108. Tran, C.T., Zhang, M., Andreae, P., Xue, B., Bui, L.T.: Multiple imputation and ensemble learning for classification with incomplete data. In: Intelligent and Evolutionary Systems, pp. 401–415. Springer, ??? (2017)
109. Oehmcke, S., Zielinski, O., Kramer, O.: knn ensembles with penalized dtw for multivariate time series imputation. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2774–2781 (2016). IEEE
110. Re, M., Valentini, G.: Ensemble methods. Advances in machine learning and data mining for astronomy, 563–593 (2012)
111. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning **36**(1-2), 105–139 (1999)
112. Adeniran, A.A., Adebayo, A.R., Salami, H.O., Yahaya, M.O., Abdulraheem, A.: A competitive ensemble model for permeability prediction in heterogeneous oil and gas reservoirs. Applied Computing and Geosciences **1**, 100004 (2019)
113. Whitehead, M., Yaeger, L.: Sentiment mining using ensemble classification models. In: Innovations and Advances in Computer Sciences and Engineering, pp. 509–514. Springer, ??? (2010)
114. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and systems magazine **6**(3), 21–45 (2006)
115. Friedman, J.H., Popescu, B.E., *et al.*: Importance sampled learning ensembles. Journal of Machine Learning Research **94305**, 1–32 (2003)
116. Ponti Jr, M.P.: Combining classifiers: from the creation of ensembles to the decision fusion. In: 2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials, pp. 1–10 (2011). IEEE
117. Bühlmann, P., Hothorn, T., *et al.*: Boosting algorithms: Regularization, prediction and model fitting. Statistical Science **22**(4), 477–505 (2007)
118. Dietterich, T.G., *et al.*: Ensemble learning. The handbook of brain theory and neural networks **2**, 110–125 (2002)
119. Chen, Y., Wong, M.-L., Li, H.: Applying ant colony optimization to configuring stacking ensembles for data mining. Expert Systems with Applications **41**(6), 2688–2702 (2014)
120. Aggarwal, C.C.: Data Classification: Algorithms and Applications. CRC press, ??? (2014)
121. Dzeroski, S., Zenko, B.: Is combining classifiers better than selecting the best one? In: ICML, vol. 2002, pp. 123–30 (2002). Citeseer
122. Khan, S.S., Ahmad, A., Mihailidis, A.: Bootstrapping and multiple imputation ensemble approaches for classification problems. Journal of Intelligent & Fuzzy Systems **37**(6), 7769–7783 (2019)
123. Aleryani, A., Wang, W., De La Iglesia, B.: Multiple imputation ensembles (mie) for dealing with missing data. SN Computer Science **1**, 1–20 (2020)
124. Wang, P., Chen, X.: Three-way ensemble clustering for incomplete data. IEEE Access **8**, 91855–91864 (2020)
125. Madhu, G., Bharadwaj, B.L., Nagachandrika, G., Vardhan, K.S.: A novel algorithm for missing data imputation on machine learning. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 173–177 (2019). IEEE
126. Lu, C.-B., Mei, Y.: An imputation method for missing data based on an extreme learning machine auto-encoder. IEEE Access **6**, 52930–52935 (2018)
127. Zhu, B., He, C., Liatsis, P.: A robust missing value imputation method for noisy data. Applied Intelligence **36**(1), 61–74 (2012)
128. Rieger, A., Hothorn, T., Strobl, C.: Random forests with missing values in the covariates (2010)
129. Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., Ning, G.: Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recognition **69**, 52–60 (2017)
130. Ali, N.A., Omer, Z.M.: Improving accuracy of missing data imputation in data mining. Kurdistan Journal of Applied Research **2**(3), 66–73 (2017)
131. Pan, R., Yang, T., Cao, J., Lu, K., Zhang, Z.: Missing data imputation by k nearest neighbours based on grey relational structure and mutual information. Applied Intelligence **43**(3), 614–632 (2015)
132. Dzulkalnine, M.F., Sallehuddin, R.: Missing data imputation with fuzzy feature selection for diabetes dataset. SN Applied Sciences **1**(4), 362 (2019)
133. Tsai, C.-F., Li, M.-L., Lin, W.-C.: A class center based approach for missing value imputation. Knowledge-Based Systems **151**, 124–135 (2018)
134. Ngouna, R.H., Ratolojanahary, R., Medjaher, K., Dauriac, F., Sebilo, M., Junca-Bourié, J.: A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values. Engineering Applications of Artificial Intelligence **95**, 103822 (2020)
135. Raja, P., Sasirekha, K., Thangavel, K.: A novel fuzzy rough clustering parameter-based missing value imputation. Neural Computing and Applications, 1–18 (2019)
136. Veras, M.B., Mesquita, D.P., Mattos, C.L., Gomes, J.P.: A sparse linear regression model for incomplete datasets. Pattern Analysis and Applications, 1–11 (2019)
137. Kim, T., Ko, W., Kim, J.: Analysis and impact evaluation of missing data imputation in day-ahead pv generation forecasting. Applied Sciences **9**(1), 204 (2019)
138. Yen, N.Y., Chang, J.-W., Liao, J.-Y., Yong, Y.-M.: Analysis of interpolation algorithms for the missing values in iot time series: a case of air quality in taiwan. The Journal of Supercomputing, 1–26 (2019)

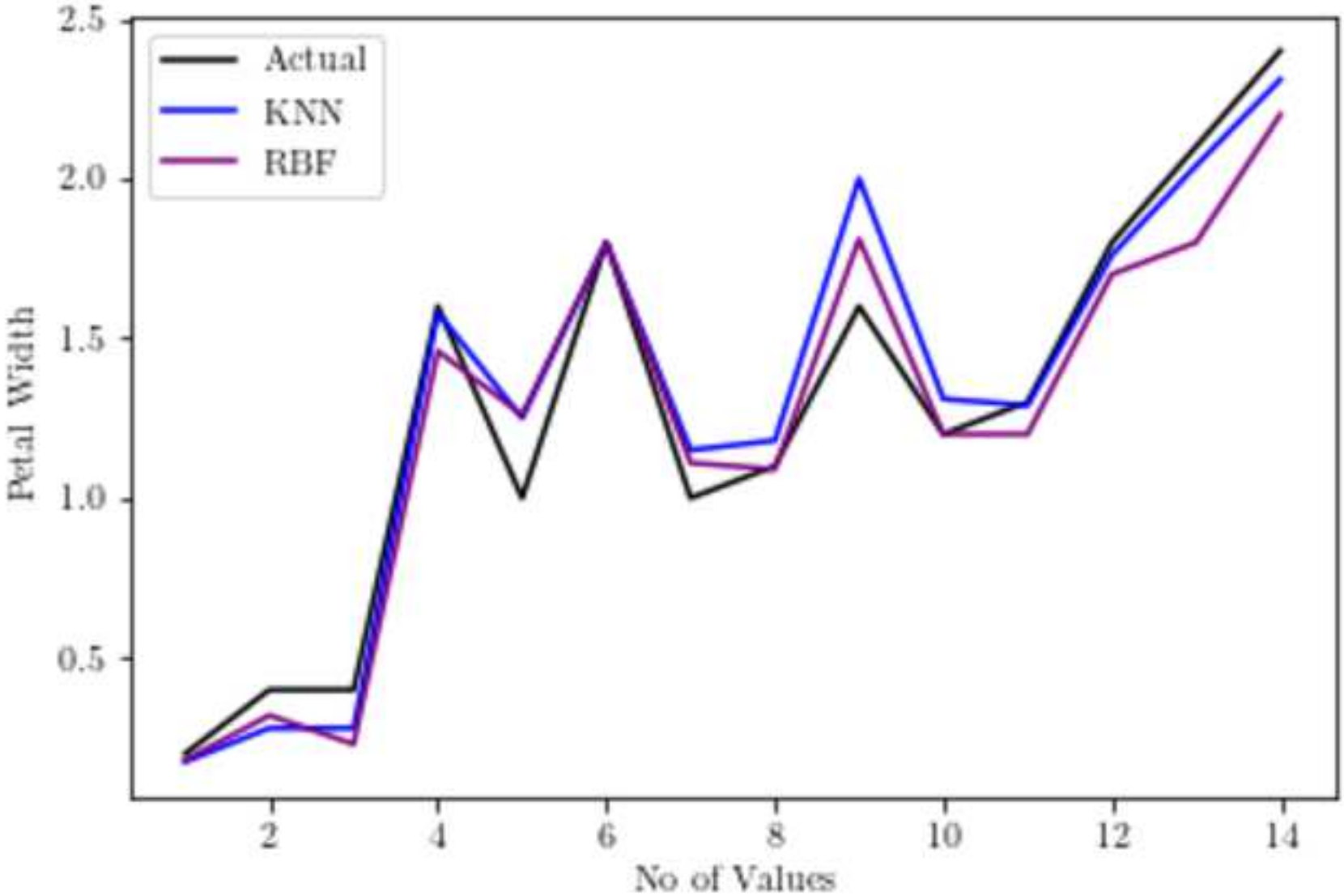
139. Ward, R.C., Axon, R.N., Gebregziabher, M.: Approaches for missing covariate data in logistic regression with mnar sensitivity analyses. Biometrical Journal (2020)
140. Beaulac, C., Rosenthal, J.S., et al.: Best: A decision tree algorithm that handles missing values. Computational Statistics, 1–26 (2020)
141. Hung, C.-Y., Jiang, B.C., Wang, C.-C.: Evaluating machine learning classification using sorted missing percentage technique based on missing data. Applied Sciences **10**(14), 4920 (2020)
142. Yang, S., Berdine, G.: The receiver operating characteristic (roc) curve. The Southwest Respiratory and Critical Care Chronicles **5**(19), 34–36 (2017)
143. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
144. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development **7**(3), 1247–1250 (2014)
145. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research **30**(1), 79–82 (2005)
146. Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C.: Semi-parametric optimization for missing data imputation. Applied Intelligence **27**(1), 79–88 (2007)
147. Deb, R., Liew, A.W.-C.: Missing value imputation for the analysis of incomplete traffic accident data. Information sciences **339**, 274–289 (2016)
148. Purwar, A., Singh, S.K.: Empirical evaluation of algorithms to impute missing values for financial dataset. In: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 652–656 (2014). IEEE
149. Willmott, C.J., Matsuura, K., Robeson, S.M.: Ambiguities inherent in sums-of-squares-based error statistics. Atmospheric Environment **43**(3), 749–752 (2009)
150. Hajian-Tilaki, K.: Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine **4**(2), 627 (2013)
151. Lobo, J.M., Jiménez-Valverde, A., Real, R.: Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography **17**(2), 145–151 (2008)
152. Son, Y., Kim, W.: Missing value imputation in stature estimation by learning algorithms using anthropometric data: A comparative study. Applied Sciences **10**(14), 5020 (2020)
153. De Silva, A.P., De Livera, A.M., Lee, K.J., Moreno-Betancur, M., Simpson, J.A.: Multiple imputation methods for handling missing values in longitudinal studies with sampling weights: Comparison of methods implemented in stata. Biometrical Journal (2020)
154. Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence **33**(10), 913–933 (2019)
155. Pu, Q., Yoo, E.-H.: Ground pm2. 5 prediction using imputed maiac aod with uncertainty quantification. Environmental Pollution, 116574 (2021)
156. Chan, R.K.C., Lim, J.M.-Y., Parthiban, R.: A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system. Expert Systems with Applications **171**, 114573 (2021)
157. Alsaber, A.R., Pan, J., Al-Hurban, A.: Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of kuwait environmental data (2012 to 2018). International Journal of Environmental Research and Public Health **18**(3), 1333 (2021)
158. Zhang, T., Zhang, D.-g., Yan, H.-r., Qiu, J.-n., Gao, J.-x.: A new method of data missing estimation with fnn-based tensor heterogeneous ensemble learning for internet of vehicle. Neurocomputing **420**, 98–110 (2021)
159. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z.: Missing value estimation for mixed-attribute data sets. IEEE Transactions on Knowledge and Data Engineering **23**(1), 110–121 (2010)
160. Slavakis, K., Giannakis, G.B., Mateos, G.: Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. IEEE Signal Processing Magazine **31**(5), 18–31 (2014)
161. Le, N.T., Benjapolakul, W.: A data imputation model in phasor measurement units based on bagged averaging of multiple linear regression. IEEE Access **6**, 39324–39333 (2018)
162. Krause, R.W., Huisman, M., Steglich, C., Sniiders, T.A.: Missing network data a comparison of different imputation methods. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 159–163 (2018). IEEE
163. Xu, X., Chong, W., Li, S., Arabo, A., Xiao, J.: Miaec: Missing data imputation based on the evidence chain. IEEE Access **6**, 12983–12992 (2018)
164. Osman, M.S., Abu-Mahfouz, A.M., Page, P.R.: A survey on data imputation techniques: Water distribution system as a use case. IEEE Access **6**, 63279–63291 (2018)
165. Pristyanto, Y., Pratama, I.: Missing values estimation on multivariate dataset: Comparison of three type methods approach. In: 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 342–347 (2019). IEEE
166. McMahon, P., Zhang, T., Dwight, R.A.: Approaches to dealing with missing data in railway asset management. IEEE Access **8**, 48177–48194 (2020)
167. Zhao, J., Nie, Y., Ni, S., Sun, X.: Traffic data imputation and prediction: An efficient realization of deep learning. IEEE Access **8**, 46713–46722 (2020)
168. Fazakis, N., Kostopoulos, G., Kotsiantis, S., Mporas, I.: Iterative robust semi-supervised missing data imputation. IEEE Access **8**, 90555–90569 (2020)
169. Aidos, H., Tomás, P.: Neighborhood-aware autoencoder for missing value imputation. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 1542–1546 (2021). IEEE
170. Li, D., Zhang, H., Li, T., Bouras, A., Yu, X., Wang, T.: Hybrid missing value imputation algorithms using fuzzy c-means and vaguely quantified rough set. IEEE Transactions on Fuzzy Systems (2021)
171. Hossain, T., Ahad, M., Rahman, A., Inoue, S.: A method for sensor-based activity recognition in missing data scenario. Sensors **20**(14), 3811 (2020)

172. Zhu, C., Chen, C., Zhou, R., Wei, L., Zhang, X.: A new multi-view learning machine with incomplete data. Pattern Analysis and Applications, 1–32 (2020)
173. Rani, P., Kumar, R., Jain, A.: Multistage model for accurate prediction of missing values using imputation methods in heart disease dataset. In: Innovative Data Communication Technologies and Application, pp. 637–653. Springer, ??? (2021)
174. Dong, W., Gao, S., Yang, X., Yu, H.: An exploration of online missing value imputation in non-stationary data stream. SN Computer Science **2**(2), 1–11 (2021)
175. Rani, S., Solanki, A.: Data imputation in wireless sensor network using deep learning techniques. In: Data Analytics and Management, pp. 579–594. Springer, ??? (2021)
176. Kabir, G., Tesfamariam, S., Hemsing, J., Sadiq, R.: Handling incomplete and missing data in water network database using imputation methods. Sustainable and Resilient Infrastructure **5**(6), 365–377 (2020)
177. Wahl, S., Boulesteix, A.-L., Zierer, A., Thorand, B., Van De Wiel, M.A.: Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC medical research methodology **16**(1), 1–18 (2016)
178. Kumar, N., Hoque, M., Sugimoto, M.: Kernel weighted least square approach for imputing missing values of metabolomics data (2021)
179. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)

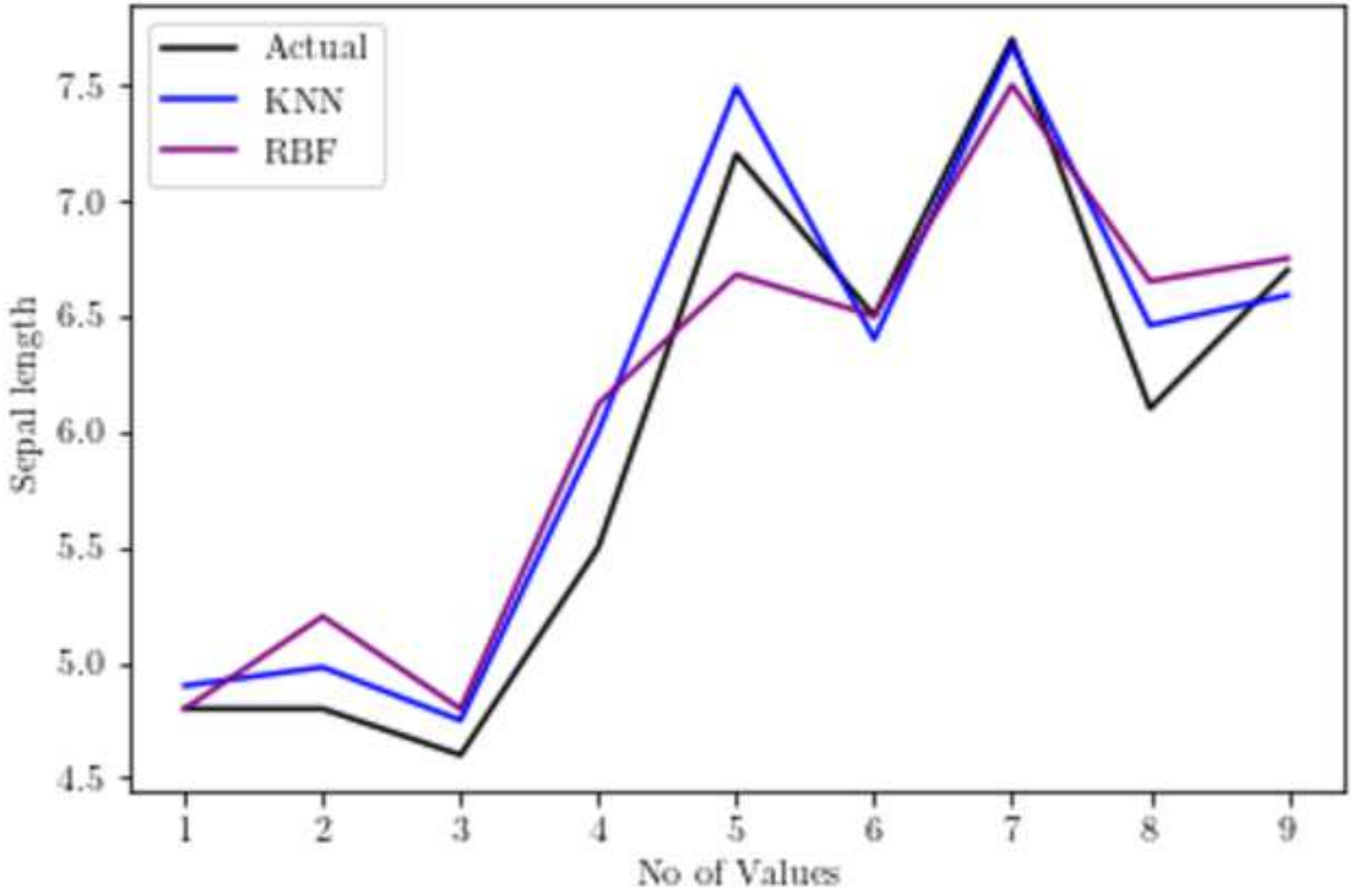
Figure


Actual
KNN
RBF
IdFanVibrations
No of Values
2.5
2.0
1.5
1.0
0.5
0
5
10
15
20

Figure


Actual
KNN
RBF
2.5
2.0
1.5
1.0
0.5
Petal Width
2
4
6
8
10
12
14
No of Values

Figure


Actual
KNN
RBF
7.5
7.0
6.5
6.0
5.5
5.0
4.5
Sepal length
1
2
3
4
5
6
7
8
9
No of Values