International Journal of Innovative Technology and Interdisciplinary Sciences
www.IJITIS.org
ISSN:2613-7305
Volume 5, Issue 3, pp. 971-1005, 2022
DOI: https://doi.org/10.15157/IJITIS.2022.5.3.971-1005
Received July 28, 2022; Accepted September 8, 2022

A Review of Missing Data Handling Techniques for Machine Learning

Luke Oluwaseye Joel [*a1], **Wesley Doorsamy** [b1], **Babu Sena Paul** [c1]

[1]Institute for Intelligent Systems
University of Johannesburg, Johannesburg, South Africa

[*a]oluwaseyejoel@gmail.com; [b]wdoorsamy@uj.ac.za; [c]bspaul@uj.ac.za

ABSTRACT

Real-world data are commonly known to contain missing values, and consequently affect the performance of most machine learning algorithms adversely when employed on such datasets. Precisely, missing values are among the various challenges occurring in real-world data. Since the accuracy and efficiency of machine learning models depend on the quality of the data used, there is a need for data analysts and researchers working with data, to seek out some relevant techniques that can be used to handle these inescapable missing values. This paper reviews some state-of-art practices obtained in the literature for handling missing data problems for machine learning. It lists some evaluation metrics used in measuring the performance of these techniques. This study tries to put these techniques and evaluation metrics in clear terms, followed by some mathematical equations. Furthermore, some recommendations to consider when dealing with missing data handling techniques were provided.

Keywords: Machine learning; Data; Missing Data; Techniques; Classification model.

1. INTRODUCTION

In general terms, missing values, also referred to as missing data, are the values not stored or captured for some variables of interest [1]. In using machine learning (ML) algorithms for predictive models, it is common to find some missing values in the dataset given or obtained. The missing values considered in this paper do not include a scenario where the number of observed cases of a feature(s) is significantly lower than the number of observed cases in other features in the dataset. This type of scenario is referred to as an imbalanced class or data. The category of data considered in this study is a balanced dataset with missing values in some features of the data. These missing values are also referred to as item non-response in survey data [2, 3]. Figure 1 is an example of such type of missing data which contains a dataset for employees with 1000 cases and 6 features [4].

It is to be noted that, among missing data handling methods, deletion methods (either listwise or pairwise deletion) are not the only option, and sometimes not the best option. This is because they can introduce bias in the estimates if the missing values are not missing completely at random (MCAR) and can reduce statistical power [1, 5]. Additionally, since missing values carry relevant information for the prediction task, it is good to impute the missing values rather than deleting them. Furthermore, some decision tree methods such as random forests have a built-in method of handling missing values, treating the missing values as an attribute (see Section 3 for that).

In one way or the other, these missing values must be handled to help the ML algorithms perform optimally. It should be noted that the missing data handling

considered in this study is within the scope of supervised learning methods. Handling missing values is one of the processes in data preparation in ML. Data preparation or pre-processing takes a considerable amount of time in modeling and prediction process because real-world data could contain missing values, outliers, conflicting data or noise. Hence, the need for data pre-processing to resolve these data issues to ensure that the data used for modeling, and predictions are of good quality. The accuracy and efficiency of ML models depend on the quality of data used for the analysis.

Some of the factors responsible for these missing values could be due to the way the data is captured or some malfunctions in the equipment. It could also be because of changes in experimental design during the process of collecting data, or the method used in merging of several datasets. Refusal of the respondents to respond to certain questions, or some variables not being measured due to some damages in the specimen used could also be a factor. The need to handle these missing values becomes highly important in ML models when the variables containing them contribute substantially to the performance of the ML algorithms. Hence, the demand to find a way of handling such missing values.

The rest of this paper is organized as follows: Section 2 focuses on the three types of missing data mechanism. Section 3 is on the explanation of some techniques for handling missing data. Section 4 reviews some evaluation metrics for measuring the performance of the missing data techniques. Section 5, outlines some discussions and recommendations observed while writing the paper.

2. MISSING DATA MECHANISM

Data missingness is generally categorized into three types [6], which are missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR), which is also called missing not at random (MNAR). Let X be the data matrix (e.g., Figure 1) that contains both the observed and missing values, X_{obs} be the set of observed values, X_{miss} be the set of missing values, R is the indicator matrix for the missing values, with

$$R_{ij} = \begin{cases} 1, & if\ X_{ij}\ is\ missing \\ 0, & if\ \ X_{ij}\ is\ observed \end{cases} \tag{1}$$

where$i$ represents the i^{th} case and j represents the j^{th} feature, and φ is a vector of some unknown parameters of the missing data model that relates the data matrix, X, to the indicator matrix,R. Brief explanation of these three categories is given below concerning these defined variables.

2.1 MCAR

In missing completely at random (MCAR), the probability of the missing data in a case does not depend on the observed or known values nor on the missing values of that case. That is,

$$Prob(R = 1|X, \varphi) = Prob(R = 1|\varphi)\ \ for\ \ all\ \ \ X, \varphi \tag{2}$$

Missingness obtained when data are missing by design, equipment failure, or when samples are lost in transit is classified as MCAR.

For example: if a missing value for an individual in the salary variable of the dataset, in Figure 1, does not depend on whether the person is a male or female, nor on the team the person belongs to, nor on whether the person is categorized as senior management or not, nor on the bonus percentage of the person, nor his/her first name, nor on the fact that information is missing from the person's record in any of the variables in the dataset but it depends on other parameters that are not captured in the dataset.

	First Name	Gender	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	97308	6.945	TRUE	Marketing
1	Thomas	Male	61933	NaN	TRUE	NaN
2	Maria	Female	130590	11.858	FALSE	Finance
3	Jerry	Male	NaN	9.34	TRUE	Finance
4	Larry	Male	101004	1.389	TRUE	Client Services
5	Dennis	n.a.	115163	10.125	FALSE	Legal
6	Ruby	Female	65476	10.012	TRUE	Product
7	NaN	Female	45906	11.598	NaN	Finance
8	Angela	NaN	NaN	18.523	TRUE	Engineering
9	Frances	Female	139852	7.524	TRUE	Business Development
10	Louise	Female	63241	15.132	TRUE	NaN
11	Julie	Female	102508	12.637	TRUE	Legal
12	Brandon	Male	112807	17.492	TRUE	Human Resources
13	Gary	Male	109831	5.831	FALSE	Sales
14	Kimberly	Female	41426	NaN	TRUE	Finance
15	Lillian	NaN	59414	1.256	FALSE	Product
16	Jeremy	Male	90370	7.369	FALSE	Human Resources
17	Shawn	Male	111737	6.414	FALSE	na
18	Diana	Female	132940	19.082	FALSE	Client Services
19	Donna	Female	81014	1.894	FALSE	Product
20	Lois	NaN	64714	4.934	TRUE	Legal
21	Matthew	Male	100612	13.645	FALSE	Marketing
22	Joshua	NaN	90816	18.816	TRUE	Client Services
23	NaN	Male	125792	5.042	NaN	NaN
24	John	Male	97950	13.873	FALSE	Client Services
25	NaN	Male	37076	18.576	NaN	Client Services
26	Craig	Male	37598	7.757	TRUE	Marketing
27	Scott	NaN	122367	5.218	FALSE	Legal
28	Terry	Male	124008	13.464	TRUE	Client Services
29	Benjamin	Male	79529	7.008	TRUE	Legal

Figure 1. Example of a Dataset with Missing Values [4].

2.2 MAR

Data that are missing at random (MAR) when the probability of the missing data in a case does depend on the observed values but not on the missing data of that case. That is,

$$Prob(R = 1|X, \varphi) = Prob(R = 1|X_{obs}, \varphi) \quad forall X_{miss}, \varphi \tag{3}$$

It should be noted that both MAR and MCAR are said to be ignorable missing data mechanisms [6, 7], due to the randomness contained in their missing data. MAR is a weaker assumption than MCAR. That is, MAR is more general, and a more realistic case of MCAR. For example: if a missing value for an individual in the salary variable of the dataset, in Figure 1, depends on the fact that the person is a female, because of some permissions of leave given to the female employees in the company which the male counterparts are not entitled to.

2.3 NMAR

These are missing data that are neither MCAR nor MAR. That is, the probability of missingness in a case depends on either the missing values or on both the missing and observed values for that case. That is,

$$Prob(R = 1|X, \varphi) = Prob(R = 1|X_{miss}, \varphi) \quad for\ all\ \ X_{miss}, \varphi \tag{4}$$

or

$$Prob(R = 1|X, \varphi) = Prob(R = 1|X_{obs}, X_{obs}, \varphi) \quad for \quad all \quad X, \varphi \tag{5}$$

The case of missing data in NMAR seems more problematic, and an unbiased estimation of these missing data can only be obtained by modeling of these missing data. NMAR is also known as a non-ignorable missing data mechanism [6, 7], that is the missingness is not random. For example: if a missing value for an individual in the team variable of the dataset, in Figure 1, depends on the fact that such person's first name is missing from the dataset or that both the first name is missing, and the person is a female who probably has gone on maternity leave.

3. TECHNIQUES FOR HANDLING MISSING DATA

This section explains the various techniques that have been used in handling missing data in ML algorithms as well as relevant literature in which they have been used. The list (see Figure 2) is not exhaustive, and this field of missing data handling is still evolving. Generally, there are two notable methods of handling missing values in ML algorithms: by deletion or imputations. However, as mentioned in the introduction of this paper, some decision tree methods have a built-in method of treating the missing values as an attribute. In the deletion techniques, the data scientist either delete the missing values by listwise or pairwise deletion. The other way is by using imputation techniques, where the missing values are replaced with some other values according to the various methods that are used. These imputation methods could be classified into single, multiple, model-based, machine learning, or optimization algorithm imputations (see Figure 2). It is good to note that some of the model-based imputation methods are also single-imputation, such as regression, K-Nearest Neighbor and hot-deck imputations. LOCF, in Figure 2, means last observation carried forward and NOCB means next observation carried backward.

Let X denote a dataset with several observations as given below

$$X = \begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1j} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2j} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & X_{i3} & \cdots & X_{ij} \end{pmatrix} \tag{6}$$

The value i is the number of cases and j is the number of independent attributes. Hence, X_{ij} is the value of the i^{th} case with the j^{th} feature. This can also be represented in terms of row and column. So, X_{ij} is the value at the i^{th} row and j^{th}column.

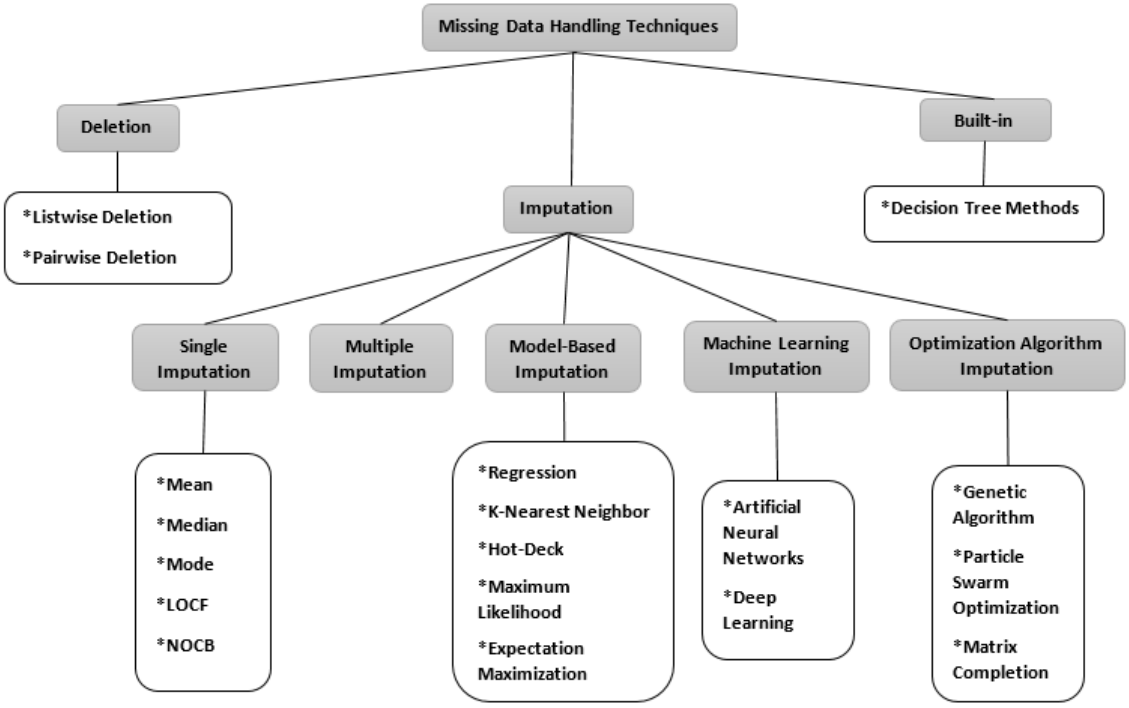

Figure 2. Types of Techniques for Handling Missing Data

3.1 Listwise Deletion

This is the removal of all cases with missing values in at least one of the independent features. It is also called complete-case analysis method [1, 5, 8]. For example, if X_{1k} is a missing value in the 1st case and k^{th} feature, then all the data in the 1st case, $X_{11}, \ldots, X_{1j}$ will be removed from the dataset. An advantage of the listwise deletion is its convenience. For data that are MCAR, this method produces unbiased estimates of means and variances [9], otherwise the estimates are biased [5]. However, the disadvantage of listwise deletion is reduction in the effective sample size which, in turn causes a reduction in statistical power.

3.2 Pairwise Deletion

Pairwise deletion [1], also known as available case method [6, 10], removes all cases containing missing values in the independent features as the need arises. For example, if X_{1k} is a missing value in the 1st case and k^{th} feature, then 1st case will be removed from the dataset when conducting analysis that involve the k^{th} feature. Unlike the listwise deletion that removes all the cases containing the missing values before the analysis, pairwise deletion removes the cases containing the missing values only when the affected features are to be used in analyses [11]. Pairwise deletion allows you to use more of your data than the listwise deletion. One disadvantage of using pairwise deletion is that the correlation matrix may not be positive definite [1]. It could produce a situation where one or more variables are linear combination of some other variables in the dataset. Another disadvantage is the lack of consistency in its analyses because each analysis will make use of different subsets of the dataset which will give different data sample sizes.

3.3 Mean Imputation

The mean imputation method [5] calculates the average values of all the non-missing values of each independent feature and imputes it for each feature's missing values [9]. Let X_{hk} be the missing value for the h^{th} case, where $h \leq i$, and in the k^{th} feature, then the mean value to be imputed will be

$$X_{hk} = \frac{\sum_{q \in I_k} X_{hq}}{n_k}, \tag{7}$$

where the value, I_k, is the set of the independent feature indices that are not missing and n_k is the total number of cases in which the k^{th} value is not missing. Although this method is easy and simple to implement, it works only if the feature considered is not nominal and the missingness assumption is MCAR [7]. Some disadvantages of the mean imputation are overestimation of the sample size, underestimation of the variance, and correlation could be negatively biased [5, 12]. Despite these disadvantages, the mean imputation still performed better than other imputation techniques in some cases. For example, in [13], the following univariate imputation methods: Mean, Median, Last Observation Carried Forward, Kalman Filter, Random, and Markov imputations were used to handle missing data in short-term monitoring (less than 24hours) of air pollutants. And the study found out that the mean imputation with two others - Random and Markov's methods outperformed the remaining univariate imputation methods recording the lowest error and highest R^2 (coefficient of determination) values for all the four periods of missingness considered.

3.4 Median Imputation

This method [13] substitutes the middle value of the non-missing independent features for each feature's missing values. The median imputation is preferred over the mean when outliers are present in the dataset [14]. Let X_{hk} be the missing value for the h^{th} case, where $h \leq i$, and in the k^{th} feature. To find the median imputation, first, the non-missing values, $X_{1j}, X_{2j}, \ldots X_{ij}$ in the feature which contains the missing value (but excluding the missing value) will have to be ordered from the lowest to the highest, and the middle value will be selected. The imputed value after sorting the data in the specific feature containing the missing value in ascending order will be

$$X_{hk} = \left(\frac{n_k+1}{2}\right)^{th} value \tag{8}$$

for odd numbers, and

$$X_{hk} = \frac{\left(\left(\frac{n_k}{2}\right)^{th} value + \left(\frac{n_k}{2}+1\right)^{th} value\right)}{2} \tag{9}$$

for even numbers. n_k is the total number of cases in which the k^{th} value is not missing. However, for a feature containing integer values, Equation (5) must be rounded up or down depending on the resulted value from the calculations. In [14], the author compared four imputation methods namely mean, median, linear regression, standard deviation, for handling missing data with different percentages of missingness and found out that the median imputation outperformed the remaining three methods.

3.5 Mode Imputation

In this method [8], the missing values of a feature are replaced by the most frequent non-missing value of that feature. Let X_{hk} be a missing value for the h^{th} case, where $h \leq i$ and in the k^{th} feature. Then, the most frequently occurring value among the non-missing values of that feature, $X_{1k}, X_{2k}, \ldots X_{ik}$, is used to replace all the missing values of that feature. The mode imputation is common for categorical features, though it can also be used for numeric features. Although this method is easy and fast to use but it could change the statistical nature of the data [15]. It is highly biased when applied to numerical data. Another challenge in using this method is when this mode value is not unique. In this case, the researcher must decide whether to use the first occurring mode value or the others. In [8], the author compared the performance of different ML classification algorithms based on some methods used in handling the missing data. Although the study focusses on the performance of the various ML classification algorithms used, but the result shown in the paper gives insight into the performance of the techniques used in dealing the missing data. Out of six methods used to handle the numeric data, mode imputation is one of the two that outperformed the remaining four methods. While for categorical data, only two methods were used to handle the missing value: listwise deletion and mode imputation. The later performed better than the former.

3.6 Last Observation Carried Forward

This method [1, 13, 16], Last Observation Carried Forward (LOCF), replaces the missing values in a feature with the last observed value of that feature. This is a common imputation method for a longitudinal or time-series dataset, in which variables are repeatedly measured over a series of time-points [16]. Let X_{hk} be the missing value for the h^{th} case, where $h \leq i$, and in the k^{th} feature. Then, the X_{hk} will be replaced by X_{hk-1}, if it exists. However, if X_{hk-1} does not exist, that is, it is also a missing value, then X_{hk-2} will be used to replace the missing values - X_{hk-1} and X_{hk}. This forms the pattern of LOCF method. Although this method is easy to understand and apply but it produces a biased estimation and underestimates the variability of the estimated result [16]. The following studies have criticized the use of LOCF method for handling missing values [16, 17].

3.7 Multiple Imputation

The multiple imputation (MI) method [18, 19, 20, 21, 22] produces a range of m possible values, $m > 1$, called the imputed data, from the existing data, which is then analysed using some standard statistical methods to obtain the most suitable set of values for the missing data. The MI method contains three (3) steps namely:

- **The Imputation step:** This is where $m, m > 1$, copies of the missing data are created or produced using appropriate model that incorporates a random variation and fits the distribution assumptions of the data. One of the popular methods is to use linear regression imputation using

$$X_{miss} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + sE \tag{10}$$

where X_1, X_2, X_3 are fully observed variables, X_{miss}, contains the missing data, E is a random value drawn from a standard normal distribution with a mean and standard deviation of 0 and 1, respectively, and s is the error term estimates of the standard deviation.

- **The analysis step:** In this step, some statistical analyses are carried out on the m imputed dataset produced in the imputation step.

- **The pooling step:** This step combines the m sets of estimation results into a single set of results. This is done using Rubin's rule [6]. The variance for the pooling can be calculated as:

For within-imputation variance

$$Var_{with} = \frac{1}{m}\sum_{i=1}^{m} s_i^2 \tag{11}$$

and for between-imputation variance

$$Var_{btw} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\beta_i - \overline{\beta}\right)^2. \tag{12}$$

The total variance could be calculated as

$$Var_{total} = Var_{with} + Var_{btw} + \frac{Var_{btw}}{m} \tag{13}$$

which gives the following formula

$$Var_{total} = \frac{1}{m}\sum_{i=1}^{m} s_i^2 + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{i=1}^{m}\left(\beta_i - \overline{\beta}\right)^2 \tag{14}$$

and the estimated standard errors will be

$$Std_{error} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} s_i^2 + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{i=1}^{m}\left(\beta_i - \overline{\beta}\right)^2} \tag{15}$$

where s_i is the standard error in the i^{th} dataset, β_i is the estimated parameter for i^{th} dataset of m samples, and $\overline{\beta}$ is the mean of β_i.

The MI method helps to restore the natural variability of the missing values, produces valid statistical inference, and generates appropriate results in the presence of a high volume of missing values [18, 19, 22]. It is flexible and can be used for MCAR, MAR and MNAR data missing mechanism. In [19], the author described the context and situations for which MI method is appropriate. The author emphasized that the goal of MI is to provide statistically valid inference in a situation where the end-users and database constructors are distinct entities [20] and where the missing data could not be traced to any admissible reason. In [18], the author provided the theory and the implementation of the MI method using Alzheimer's disease as a case study. The study also discussed the increasing use of MI methods in commercial and free software. Some of these software packages [18] that support and implement MI methods are, the list is

just a few: STATA [23], S-PLUS [24], AMELIA [25], IVEware [26], SPSS [27, 28], and SAS [29].

There are two most popular implementations of the MI method in the literature: (1) Multivariate imputation by chained equation (MICE) [30-34] for fully conditional specification (FCS) approach to MI and (2) AMELIA [25, 35, 36] for the joint modeling (JM) approach to MI method. In the FCS approach, the MI method imputes the missing data on a variable-by-variable basis [10, 37, 38]. In contrast, in the JM approach, the MI method imputes the missing variables based on the assumptions of multivariate normality and linearity [39, 21, 40]. These adopted assumptions in the JM approach are relaxed in the FCS approach which consequently gives the FCS approach a great deal of flexibility, appropriate coverage, popularity, and yields estimates that are generally unbiased [30, 38, 41-43]. In [44], the author compared the accuracy of four techniques for handling missing data in mental measurement questionnaires. The methods were direct deletion, mode, Hot-deck, and multiple imputation by absolute deviation imputations. The study found that the multiple imputation method performed best than other methods because the biases obtained in it are the smallest in all the proportion of missingness considered.

3.8 Regression Imputation

This is also called conditional mean imputation [7] or predictive mean imputation [9]. In this technique, missing values are replaced by predicted values a using regression model on non-missing values of the other features [7, 9]. Each missing value is taken as a target variable and the other attributes taken as independent features. Let X_{hk} be the a missing value for the i^{th} case, where $h \leq i$, and in the k^{th} feature, then the regression imputed value will be

$$X_{hk} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_3 X_{i3} \tag{16}$$

These methods are based on the assumptions that there is a linear relationship between the features. However, if the assumption does not hold true, that is, if the relationship is non-linear, this method will introduce biases into the dataset. In [45], one of the four methods used in handling missing data on crop yield dataset from the National Agricultural Statistical Survey (NASS) barley crop yield in 1997, United State of America, was regression imputation. Kernel smoothing, universal kriging, and multiple imputation methods were the other three. The result of the study showed that regression imputation and multiple imputation methods outperformed the remaining two methods.

3.9 Hot-Deck Imputation

This method uses values from among the most similar cases of non-missing data to replace the missing values [9, 46, 47]. Let X_{hk} be a missing value for the i^{th} case, where $h \leq i$, and in the k^{th} feature, then the Hot-Deck imputation is calculated using the following

$$X_{hk} = X_{pk},$$

where

$$p = arg\min_{l} \sqrt{\sum_{k \in I_{(non-missing)}} \sigma_k \left(X_{hk} - X_{lk} \right)^2} \tag{17}$$

The value,σ_k, is the standard deviation of the k^{th} non-missing feature. Although this method preserves the sample distribution [12] for the substituted missing data, it might

alter the relationship between the features. Hot-deck imputation is suitable for handling missing values where listwise deletion, mean, or median imputation will not work well [5]. Generally, this method is divided into the random hot-deck and the sequential hot-deck methods [47, 44]. Unlike regression imputation, hot-deck imputation does not depend on model fitting to impute the feature, hence it is likely not affected by model misspecification [47].

3.10 K-Nearest Neighbor Imputation

This is also known as distance-function matching [12]. This method [48] makes a random selection of values from its k nearest or closest similar cases and the one with the smallest distance is taken and used in replacing the missing value [5, 48, 49]. Let X_{hk} be a missing value for the h^{th} case, where $h \leq i$, and in the i^{th} feature, using a simple distance function, the missing value can be obtained as follows:

$$X_{hk} = X_{qk}, \quad q = \min_{p} d(X_{hk}, X_{pk}) \tag{18}$$

and $d(X_{hk}, X_{pk})$ could be

$$d(X_{hk}, X_{pk}) = \sum_{i \in KNN(X)} |X_{hk} - X_{ik}| \tag{19}$$

or

$$d(X_{hk}, X_{pk}) = \sqrt{\sum_{i \in KNN(X)} (X_{hk} - X_{ik})^2} \tag{20}$$

where K-NN (X_{hk}) is the index of the k^{th} closest cases of X_{ij} from the non-missing features. This method is different from others in that an actual value is imputed and not a constructed value as in regression imputation. When there is no prior knowledge about the data distribution, K-nearest neighbor imputation is the appropriate choice to go for. KNN imputation method is a function of the choice of distance measure used. The literature contains many different distance metrics [50] that can be used. Some KNN extensions are sequential KNN [51], interactive KNN [52, 53], and the combination of KNN and local-least square model [54-56].

In [49], the author compared some imputation methods for handling missing scores (data) in biometric fusion. The study focused on multibiometric systems, which is the fusion of multiple biometric information sources. This is because multibiometric systems perform better in recognition than uni-modal systems. The imputation techniques used for this study were K-nearest neighbor (KNN), the maximum likelihood-based, Bayesian-based, and multiple imputation (MI) methods. The experiments performed by the authors for the imputation of missing scores at a different rate of missingness showed that the KNN-based imputation method outperformed others. Also, in [57], the author compared the following imputation methods - mean, median, predictive mean matching, KNN, Bayesian linear regression, non-Bayesian linear, and random sample imputations - for handling missing numeric data. The study found that KNN imputation method performed the best.

3.11 Maximum Likelihood

This method employs all available cases of a given data to construct some parameters that would maximize the probability of those values that have been observed [7, 49]. This is accomplished by making use of a formula that gives the probability of the data as a function of both the data and the parameters to be estimated. Let X be the complete data as shown in the matrix above, with the associated probability density $f(\theta \vee X)$, where θ is the unknown parameter. If X_{obs} and X_{miss} represent the observed and missed data respectively, then $X = (X_{obs}, X_{miss})$. Hence, the objective of this method is to maximize the likelihood

$$g(\theta|\, X_{obs}) = \int_{X_{miss}} f(\,\theta|X_{obs}, X_{miss}\,)\; dX_{miss} \tag{21}$$

if X is continuous. And if X is discrete, the equation becomes

$$g(\theta|\, X_{obs}) = \sum_{X_{miss}} f(\,\theta|X_{obs}, X_{miss}\,) \tag{22}$$

To get the overall likelihood, the products of the likelihoods of all cases is taken. If there are m cases with complete data and $n - m$ cases with missing data, then the likelihood function to be maximized, to get the maximum likelihood of θ, for the whole dataset is

$$L(\theta|X_{obs}) = \prod_{i=1}^{m} f(\theta|X) \prod_{m+1}^{n} g(\theta|X_{obs}) \tag{23}$$

where $\prod$ is a repeated multiplication. For parameter under the missingness mechanism of MAR and MCAR, this method gives unbiased estimates and standard errors [7]. The maximum likelihood imputation performs better in larger datasets than in smaller samples [6] and it is sensitive to the choice of initial starting values [7].

3.12 Expectation-Maximization

This method generalises the maximum likelihood estimation to the incomplete dataset [1, 7]. It attempts to find the unknown parameters $\hat{\theta}$ that maximizes the log probability density $g(\theta \vee X_{obs})$ of the observed data. It predicts the missing data using some assumed values from the parameters [7]. Then, it updates the parameters using the predictions. The process is repeated until the sequence of parameters converges to the maximum likelihood estimates. Basically, Expectation-Maximization (EM) method has two steps [1, 7], the E-step (predicting the missing values) and the M-step (estimating the parameters). The EM algorithm is as seen in Algorithm 1.

Algorithm 1 EM Algorithm

Start with initial guess $\hat{\theta}^{(0)}$ for θ

Compute $Q(\theta, \hat{\theta}^{(t)})$

Estimate $\hat{\theta}^{(t+1)} = \arg\max_{\theta} Q(\theta, \hat{\theta}^{(t)})$

Stop when the likelihood converges

EM assumes MAR data missing mechanism and its convergence to a local maximum of the likelihood function is said to be guaranteed. However, its convergence rate depends on the fractions of missing data. Low missingness produces fast convergence, while high missingness produces slow convergence.

In [58], the author employed six imputation methods for handling the missing data in air quality, Malaysia. The imputation methods are mean, median, EM, singular value decomposition (SVD), KNN, and sequential K-nearest neighbor (SKNN) imputation methods. Using the correlation coefficient, the index of agreement and the mean absolute error as evaluation metrics, the study found out that EM is one of the three best methods for all the eight monitoring stations considered, the other two were KNN and SKNN.

3.13 Artificial Neural Network

Artificial Neural Networks (ANN) refers to a computing system inspired by how biological neural network systems, such as brain, process information [59]. ANN is also known as neural nets or neural networks or artificial neural systems. Just as the human or biological neural network contains many interconnected neurons, the ANN also consists of multiple layers of simple processing elements called artificial neurons that are interconnected to perform the function of collecting inputs and generating output [59, 60]. With the great potential of ANNs to process information with high speed in a massive parallel implementation, the use of ANN in several discipline and applications has increased over the years [61]. Particularly, in the imputation of missing values for machine learning.

In [62], the author proposed a mechanism for processing missing data by neural networks. Neuron's response in the first hidden layer was replaced by its expected value. The study noted that, this method does not require complete data for its training. That is, it trained neural network on datasets with only incomplete samples. The method was compared with other imputation methods for incomplete data in the literature, using 7 different datasets, and it gave better results than them. In [63], the author proposed an imputation method that uses an auto-encoder neural network. The auto-encoder was trained using the training data, without the missing values, to be better equipped to predict the missing values. Afterwards, the trained autoencoder was used to predict missing values, using the training data with missing values, with the idea that a good choice for replacing each missing value will be the one that can reconstruct itself by means of the auto-encoder. The method was compared with eight other imputation methods using fourteen different datasets and eight classification techniques. The results [63] showed that the performance of the proposed method is noticeably better than other methods for higher missing rates.

3.14 Deep Learning

Deep Learning (DL) [64, 65] is a subset of machine learning that has several layers (excluding the input and output layers) of ANNs that carry out the machine learning process. DL, also called deep neural network or deep nets, is an ANN with multiple hidden layers. With the multiple layers in DL, there is an advantage of performing complex tasks that often require extensive feature engineering and better self-learning capabilities. This also goes for its ability to handle missing values [66-71] in datasets.

In [66], the author used a DL method, with 15 hidden layers using Rectified Linear Unit as the activation function, for the imputation of missing data in attention-deficit/hyperactivity disorder (ADHD). The imputed datasets were used to distinguish youths with ADHD from those without it. A dataset of 1220 youths recruited in

Northern Taiwan was employed for the study with 799 youths having ADHD and 421 without ADHD. The performance of the DL method used was evaluated using support vector machine classifier, and the study found out that the result of the classifier on the imputed dataset (which is 89% accuracy) is the same as the result of the classifier on the original dataset (without missing values). This shows that the DL imputation method does not introduce any bias toward the data when imputing the missing values.

In [67], the author proposed an imputation method based on DL for imputing missing traffic data. Specifically, the authors used denoising stacked autoencoder (DSAE) for the DL architecture which consist of a denoising autoencoder (DAE) and stacked autoencoder (SAE). The study treated the traffic data (including both the observed and missing values) as a single data and conducted the complete data restoration using this method. Data from Caltrans Performance Measurement System (PeMS) was used and the results from the study showed that, at different missing rate, the errors were kept at a stable level.

In [68], the authors proposed the generative adversarial multiple imputation network (GAMIN) for highly missing data. That is, for 80% and above missing data. The generative adversarial network (GAN) [69] is a popular DL approach for missing data imputations for large datasets. However, among other changes done in GAMIN, the authors incorporated the unconditional generator directly into the imputation process, used a different confidence prediction method, a new loss function was used to train GAMIN and the mask of missing data itself was utilized instead of the mask generation. The results of the study, for both modified National Institute of Standards and Technology (MNIST) and CelebFaces Attributes (CelebA) Datasets, demonstrated the better performance of the proposed method as compared to the method used in [72, 73].

3.15 Genetic Algorithm

Genetic Algorithm (GA) is a population-based optimization algorithm, inspired by the biological evolution process [74, 75]. GA, proposed by J.H. Holland [76, 77], has the following basic elements, namely, chromosome representation, fitness selection, and biological-inspired operators (which are selection, mutation, and crossover). In form of strings, the collection of chromosomes is called a population which will be initially created randomly. This represents different points in the search space with their associated objective and fitness functions. A few of the "fittest" strings are selected into the mating pool, and the biologically inspired operators are applied on these strings to produce a new generation of strings [74, 78, 79]. This process is repeated until a satisfactory termination condition is reached, or a desired number of generations is obtained. GA imputation method has been used to address the three types (MCAR, MAR NMAR) of missing values mechanism and different types of variables in the literature [80-85].

In [80], the author proposed a method for imputing missing data based on GA and Information Gain (IG). While GA was used to generate optimal sets of missing values in the dataset, IG was used to measure the performance of each solution. The authors indicated that the imputation method is suitable for large dimensional search spaces with a higher rate of missing values. Comparing this proposed method with some existing single and multiple imputation techniques, using 5 different datasets, the study reported that the proposed method outperformed others. In [81], the authors proposed a combination of GA and Asexual Reproduction Optimization (ARO) algorithm. The combined algorithm was used to evaluated on Pima and Mammographic mass datasets, The performance was compared with other imputation techniques namely Mean, KNN, support vector machine (SVM) and it was found that it outperformed them. The combined proposed algorithm took less computational time than the basic GA.

3.16 Particle Swarm Optimization

Particle swarm optimization (PSO) algorithm is an evolutionary optimization algorithm proposed by [86]. The algorithm simulates the social behaviour of animals, such as insects, birds, and fishes, in their search pattern for food. This pattern keeps changing and updating according to the learning experiences of each member of the swarms [86, 87]. PSO has been employed in the imputation of missing data by different authors. In [88], the author proposed a missing data imputation method based on the PSO algorithm to minimize the error function obtained from the covariance matrix of the complete record (excluding the missing data) and the covariance matrix of the total records (which include the imputed data). The PSO algorithm was also used to minimize the error function derived from the determinant of these covariance matrices. The algorithm [88] stops when these two errors become acceptably small across to consecutive iterations. Similarly, in [89], the authors used the PSO to minimize these two mentioned error functions. However, the two imputation methods proposed by this study [89] involved PSO, evolving clustering method (ECM) and a modified version of autoassociative extreme learning method (MAAELM). The proposed imputation methods were tested on twelve different datasets and were compared with other hybrid imputation algorithms. One of the proposed methods, PSO + ECM, performed better than other existing algorithms in six out of the twelve datasets were used while the second one, PSO + ECM + MAAELM, performed better than other algorithms in nine out of twelve datasets used.

In [90], the missing values in the heart disease dataset were handled by an imputation method based on Fuzzy C-Means and Particle Swarm Optimization (FCMPSO). The study showed that the performance of the Decision tree, which is the classification method used, was increased after the application of FCMSPO for missing data imputation.

3.17 Matrix Completion

The matrix completion method, proposed by [91], operates with some assumptions on the data matrix to create a well-posed problem. The assumption could be that the data matrix has a maximal determinant, positive definite or low rank [92]. If it is assumed that the data matrix is low rank, the entries are correlated. Hence, the missing entries could be recovered in a convex optimization context if there is sufficient observed entries [91, 93]. The formula for this recovery method is given as [94]:

$$\begin{aligned} &Minimise \quad rank(M) \\ &subject\ to \quad P_{\Omega}(M) = P_{\Omega}(X), \end{aligned} \tag{24}$$

where M is the optimization variable, X is the matrix to be estimated, Ω is the set that contains the positions of the available observations in the matrix X, and P_{Ω} is the orthogonal projector into the subspace of matrices that vanish outside of Ω. However, for entries on Ω that are sampled uniformly at random [91], the solution for equation (24) is obtained by solving the convex relaxation equation [94]

$$\begin{aligned} &Minimise \quad \|M\|_{*} \\ &subject\ to \quad P_{\Omega}(M) = P_{\Omega}(X), \end{aligned} \tag{25}$$

$\|M\|_{*}$ denotes the nuclear norm of the matrix M. The following literature [95-99] are existing theory on matrix completion with different assumptions.

3.18 Decision Tree Method

Decision Tree [100-104], such as Random Forest [105], is a popular supervised learning algorithm that is used to build both classification and regression models. The method splits the given dataset into subsets (called decision nodes) based on the level of the significance of the attributes. The most significant predictor is known as the root node. The nodes which cannot be split further are called terminal or leaf nodes. The splits follow a set of if-else conditions, and the classification is done according to these conditions. Some attribute selective measure used in decision trees are Entropy, Information gain, Gini index, Gain Ratio, Reduction in Variance and Chi-Square. Aside from the fact that decision trees are easy to read and interpret, it also has a good way of handling missing values and outliers [101-103]. Hence, the presence of missing data and outliers have little influence on decision tree's data. Decision trees algorithms have an advantage of performing well in handling the MCAR, MAR values, and to some appreciable level in NMAR values [105]. The method performs the missing values estimation task by building a tree-like structure for each feature containing the missing value entries, and the missing values of each feature are filled by using its corresponding tree [101].

4. EVALUATION METRIC FOR MISSING DATA HANDLING TECHNIQUES

A vital step to take after employing the techniques for dealing with missing values, mentioned in Section 3, is the evaluation of the imputation methods used. There are two dimensional aspects to evaluating these missing data imputation techniques: (1) evaluation of imputation quality; and (2) evaluation using downstream prediction tasks. The two dimensions are discussed in the sub-sections below. The list is not intended to be exclusive.

4.1 Imputation Quality

The purpose of using imputation quality metrics is to compare the predicted values with the actual values to know how close the prediction is to the actual. In this section, some imputation quality metrics are selected and explained. More of these evaluation metrics can be seen in the following literature [106-109].

4.1.1 Root Mean Square Error

Most studies [13, 110, 111, 112] employed the root mean square error (RMSE) to measure the imputed and observed data difference. The RMSE represents the standard deviation of the difference, and it given as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2} \tag{26}$$

RMSE represents the quadratic mean of the imputed and observed data differences. The value of RMSE is always non-negative and a lower value is better than a higher value. One major disadvantage of RMSE is that it is sensitive to outliers.

4.1.2 Root Mean Squared Percentage Error

The root mean squared percentage error (RMSPE) is the RMSE expressed as a percentage. It is given as

$$RMSPE = \sqrt{\frac{100}{n}\sum_{i=1}^{n}\left(\frac{\left|X_i^{obs} - X_i^{imputed}\right|}{\left|X_i^{obs}\right|}\right)^2} \tag{27}$$

One prominent advantage of RMSPE is the scale-independent property [113], hence it can be used to compare forecast performance across different datasets. Smaller values of RMSPE indicates better performance of the imputation methods used.

4.1.3 Mean Squared Error

Mean Square Error (MSE) is the average squared difference between the imputed and observed values. It is given as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2 \tag{28}$$

The value of MSE is always non-negative and the values closer to zero are much preferable. MSE has a higher unit order than the error unit, this is due to the square of the error. MSE has low reliability [109], it might give different results for a different fraction of the dataset.

4.1.4 Mean Absolute Error

The average or mean of all absolute errors. It is denoted by MAE, and it is one of the top three evaluation metrics popularly used in literature [114]. It is given as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_i^{obs} - X_i^{imputed}\right| \tag{29}$$

The MAE seems to be the most intuitive evaluation metric because it takes the absolute difference between the actual and the predicted or imputed data. Also, due to the use of absolute value, there is no indication of over-performance or under-performance of the model in MAE [115]. It is robust to outliers.

4.1.5 Mean Absolute Percentage Error

The mean absolute percentage error (MAPE) is the MAE expressed as a percentage. MAPE also has an intuitive interpretation, like MAE, since percentages are easy to understand. It is one of the top three evaluation metrics popularly used in the literature [114]. It is given as

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\left|X_i^{obs} - X_i^{imputed}\right|}{\left|X_i^{obs}\right|} \tag{30}$$

Like MAE, MAPE has the advantage of the absolute value used in the formula. It is robust to outliers. However, unlike the MAE, MAPE could be undefined if the actual value is zero, grow unexpectedly large if the actual value is minimal, and it is biased if the predicted value is less than the actual value [115].

4.1.6 Mean Absolute Scaled Error

The mean absolute scaled error, MASE, is the mean absolute error of the imputed or predicted values divided by the mean absolute error of the naive prediction. MASE was proposed by [113] to be the standard evaluation measure for comparing forecast accuracy. It is given as

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i^{obs} - X_i^{imputed}\right|}{\frac{1}{n-1}\sum_{i=2}^{n}\left|X_i^{obs} - X_{i-1}^{obs}\right|} \tag{31}$$

The limitation associated with MAPE does not affect MASE. Some of the favourable properties of MASE are scale invariance, predictable behaviour as the actual value tends to 0, easy interpretation, and the ability to penalize both negative and positive predicted values as well as large and small predicted values equally [113]. MASE is robust to outliers.

4.1.7 Mean Relative Absolute Error

The mean relative absolute error (MRAE) is the mean absolute error divided by the total absolute error of the simple predictor. The MRAE expresses how large the MAE is compared to the total size of the observed data. The formula is given by

$$MRAE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|X_i^{obs} - X_i^{imputed}\right|}{\left|X_i^{obs} - \hat{X}\right|} \tag{32}$$

where $\hat{X}$ is the mean of the data. The MRAE is fairly robust to outliers but sensitive to values that are zero or almost zero. It could be undefined if the predicted value is equal to the actual value [109, 113].

4.1.8 Coefficient of Determination (R^2)

The coefficient of determination, denoted by CoD or R^2, is the square of the correlation (that is, how strong a linear relationship of two variables is) between predicted values and the actual values [13]. It is also known as the "goodness of fit". The formula is given as

$$CoD = 1 - \frac{\sum_{i=1}^{n}\left(X_i^{obs} - X_i^{imputed}\right)^2}{\sum_{i=1}^{n}\left(X_i^{obs} - \hat{X}\right)^2} \tag{33}$$

where $\hat{X}$ is the mean of the data. The values range from 0 to 1. And a higher R^2 is an indicator of a better performance than a lower R^2. The coefficient of determination metric is a widely used evaluation, but it is limited to size differences between the imputed and observed data [116].

A properties' summary of the evaluation metrics presented above can be seen in Table 1. In [117], five distance measures were used to evaluate the performance of the missing data handling techniques. In [13], five evaluation metrics namely Root Mean Square Error (RMSE), Absolute Bias, Percent Absolute Error in Means, R^2 (coefficient of determination), and Mean Absolute Error were used to evaluate the performance of the missing data handling techniques. The metric used for evaluation in [45] was the mean absolute prediction error only. While in [110], the evaluation metrics used were RMSE, unsupervised classification error (UCE), supervised classification error (SCE) and execution time. However, in [8], the author employed some classification algorithms to evaluate the performance of the techniques used in handling the missing data.

Table 1: A Summary of the Evaluation Metrics

S/N	Evaluation Metrics	Performance Value	Sensitivity to Outliers	Scale Dependency
1	RMSE	Small value indicates better performance	Yes	No
2	RMSPE	Small value indicates better performance	Yes	No
3	MSE	Small value indicates better performance	Yes	No
4	MAE	Small value indicates better performance	No	No
5	MAPE	Small value indicates better performance	No	Yes
6	MASE	Small value indicates better performance	No	No
7	MRAE	Small value indicates better performance	No	Yes
8	CoD	Value near 1 indicates better performance	No	No

4.2 Imputation Evaluation Using Downstream Prediction Tasks

The purpose of using downstream prediction tasks is to evaluate the accuracy of the imputed data after the missing data imputation process. Some downstream prediction metrics are selected and explained in this sub-section.

4.2.1 Confusion Matrix

A confusion matrix [118] (see Figure 3), also called an error matrix, is a table used to describe the performance of a classification model on the imputed data after the imputation process. The matrix is divided into four parts namely True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TP values are obtained when the classification model, after the imputations, predicts an observation belongs to a class and it does belong to that class. FP values are obtained when the model, after the imputations process, predicts an observation belongs to a class while it

does not belong to that class. FN values are obtained when the model, after the imputations, predicts an observation does not belong to a class while it belongs to that class. TN values are obtained when the model, after the imputations, predicts an observation does not belong to a class and, in truth, it does not belong to that class.

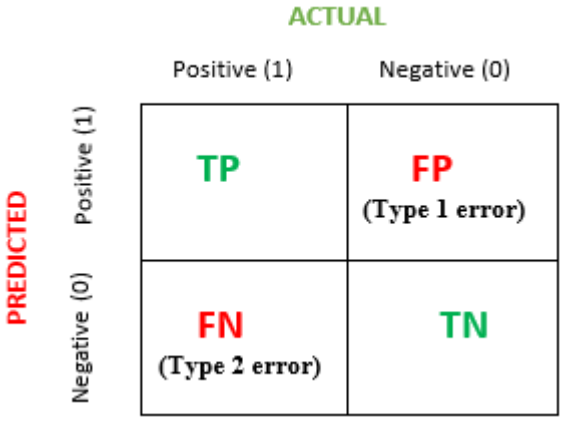

Figure 3: Confusion Matrix for Prediction on the Imputed Data

The errors incur when the classification model predicts positive based on the imputed data whereas the actual situation is negative are called type-1 errors. While the errors incur when the classification model predicts negative based on the imputed data whereas the actual situation is positive are called type-2 errors. Depending on the problem at hand, type-1 errors could be more of a concern than type-2 errors or vice versa [118].

4.2.2 Recall, Precision and Accuracy

From the confusion matrix, the following terms namely recall, precision, and accuracy can be identified and used to evaluate the performance of the imputation methods used.

Recall, also called sensitivity or True positive rate, is the percentage of the total relevant outcome correctly predicted or classified by the algorithm based on the imputed data after the imputation process. Recall should be employed when the costs of FN is high. The formula for recall is given in Equation (34).

$$Recall = \frac{TP}{TP + FN} = \frac{TruePositive}{TotalActualPositive} \tag{34}$$

Precision, which is also called Positive Predictive Value (PPV), is the fraction of positive predictions that are correct based on the imputed data after the imputation process. Precision should be employed when the costs of FP is high. The formula is given in Equation (35).

$$Precision = \frac{TP}{TP + FP} = \frac{TruePositive}{TotalPredictedPositive} \tag{35}$$

Accuracy is the fraction of the total number of predictions that were correctly predicted based on the imputed data after the imputation process. The formula for accuracy is given in Equation (36).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{CorrectPredictions}{TotalPredictions} \tag{36}$$

4.2.3 F1-score

F1-score is the harmonic mean of recall and precision. For a balance between precision and recall, F1-score is needed. This is essential for a problem seeking to achieve a good recall and precision. Obtaining a high recall means you have low precision and vice versa. Hence, the F1-score shows the predictive power of the classification model on the imputed data, after imputations on the dataset. The formula for F1-score is shown in Equation (37).

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{37}$$

4.2.4 Receiver Operating Characteristic

The Receiver Operating Characteristic (ROC) curve shows the performance of the classification model on the imputed data at all different thresholds. It is graph plotted between the TP rate ($\frac{TP}{TP+FN}$) at the y-axis and FP rate ($\frac{FP}{FP+TN}$) at the x-axis. The ROC curve is better calculated using the Area under the curve (AUC) plot. This AUC ROC plot is most popular because it checks the curve of different machine learning models using different thresholds. It is scaled variant and focuses on the quality of the model's prediction.

5. CASE STUDIES OF THE MISSING DATA HANDLING TECHNIQUES

The techniques discussed in Section 3 have been applied in different fields of study. Table 2 shows some case studies where these missing data handling techniques have been used and the different metrics employed in evaluating the performances of the imputation techniques/methods. The best imputation method reported by each of the studies is also mentioned.

Table 2: A Summary of the Missing Data Handling Techniques

S/N	Citation	Case Study	Metrics Used	Number of Imputation Methods Used	Best Imputation Method
1	Malarvizhi and Thanamani [14]	UK Census Report in 2021	Accuracy	4	Median and Standard deviation
2	Jadhav et. al. [57]	Wine Dataset, Glass Identification, Concrete Comprehensive Strength, Indian Liver Patient Dataset, Seeds Dataset	NRMSE	7	KNN
3	Ding and Ross [49]	Biometric Fusion	ROC	7	GMM-KNN
4	Le et. al. [119]	Healthcare	RMSE, Exe-	4	EM

		Dataset	cution time		
5	Cihan et. al. [120]	Veterinary Dataset	RMSE, Execution time, SCE, Accuracy, Recall, Precision, Kappa	5	missForest
6	Hadeed et. al. [13]	Air Pollutants Dataset	Absolute Bias, PAEM, R^2, RMSE, MAE	7	Markov, Random, and Mean
7	Xu et. al. [44]	Mental Measurement	AD, RMSE, ARE	4	MI
8	Krishna and Ravi [88]	Boston Housing, Forest Fire, Auto mpg, Body Fat, Pima Indians, Iris, Spectf, Wine, Banking Bankruptcy Datasets.	MAPE	2	PSO
9	Salleh and Samat [90]	Heart Disease Classification	Accuracy, Precision, ROC	4	FCMPSO
10	Gautam and Ravi [89]	Boston Housing, Forest Fire, Auto mpg, Body Fat, Pima Indians, Iris, Spectf, Wine, and Banking Bankruptcy Datasets.	MAPE	14	PSO + ECM + MAAELM
11	Shahzad et. al. [80]	Labor, Echocardiogram, Cylinder Bands, Mammographic Mass, and Colic-Horse Datasets	Accuracy, Precision, Recall, f-measure, ROC	6	GA
12	Priya and Sivaraj [83]	Microarray Classification	RMSE	4	DRAGEL
13	Priya et. al. [82]	Engineering Students Weight, Housing and Adult Datasets	Accuracy	6	GA
14	Noei and Abadeh [81]	Pima Indians and Mammographic Mass Datasets	Accuracy, ROC	5	ARO
15	Duan et. al. [67]	Traffic Data	RMSE, MAE, MRE	2	DSAE
16	Xu et. al. [121]	Adult Datasets	Accuracy	6	MIAEC
17	Sefidian and Daneshpour [122]	Iris, Wine, Glass, Haberman, Wholesale Customers, Chess,	RMSE, MAE, R^2	6	GFCMI

		and Adult Datasets.			

The following is the list of abbreviations used in Table 2 that were not in the paper before. Normalized RMSE (NRMSE); KNN imputation via Gaussian mixture model (GMM-KNN); supervised classification error (SCE); Percent Absolute Error in Means (PAEM); Absolute deviation (AD); Average relative error (ARE); Fuzzy C-Means and Particle Swarm Optimization (FCMPSO); evolving clustering method (ECM); modified auto-associative extreme learning machine (MAAELM); dynamic Bayesian genetic algorithm (DBAGEL); genetic and Asexual Reproduction Optimization (ARO); mean relative error (MRE); denoising stacked autoencoder (DSAE); missing value imputation algorithm based on the evidence chain (MIAEC); Grey based Fuzzy c-Means and Mutual Information (GFCMI).

6. DISCUSSION

In this section, some discussions and recommendations are outlined. First, it should be noted that the choice of the techniques to employ in dealing with any missing data, in each problem, depends on the nature of the missing data. Whether the missingness mechanism is either MCAR, MAR or MNAR. The listwise deletion method is unbiased for MCAR missing mechanism but gives inefficient results for all the missing data mechanism. All the techniques examined so far generate biased results for MNAR missing mechanism, and this can only be addressed by considering some additional information.

Second, it is advisable to use all the available data when performing missing data analyses. That is, the use of listwise or pairwise deletion should be highly discouraged. This is because all data provided are usable, none should be discarded, and each carries some information that could contribute meaningfully to the analysis. However, the deletion method might be used in a special situation when the data size is huge and the missing data is very small (say, less than or equal 1%). For example, listwise deletion, as mentioned before, reduces sample size and statistical power greatly, thereby increasing standard and type II error. Similarly, the use of single imputation methods should be discouraged. This is because most of the single imputation methods are biased under MCAR mechanism, and they fail to produce correct standard errors for hypothesis testing.

Third, a good and clear understanding of the different evaluation metrics' statistical properties is needed to select the appropriate measures for the performance of the different imputation methods that might be used in handling the missing data. This is important because each metric has some disadvantages that could lead to inaccurate performance measurements. Table 3 shows some advantages and disadvantages of each of the missing data handling techniques.

Table 3: Advantages and Disadvantages of Missing Data Handling Techniques

S/N	Imputation Method	Advantages	Disadvantages
1	Listwise Deletion	- It is easy to implement. - Appropriate for a small number of missing values and large dataset.	- May cause bias in the estimates. - It requires a large data sample to be used.

			- Not appropriate for a large number of missing values. - It reduces statistical power.
2	Pairwise Deletion	- It is easy to implement. - Appropriate for a small number of missing values and large dataset.	- May cause bias in the estimates - It requires a large data sample to be used. - Not appropriate for a large number of missing values. - It reduces statistical power
3	Mean	- It is easy to implement. - Appropriate for a small number of missing values.	- May cause bias in the estimates - Leads to an underestimate of the errors - May cause changes in the co-variance and variance. - Not appropriate for a large number of missing values.
4	Median	- It is easy to implement. - Appropriate for a small number of missing values.	- May cause changes in the co-variance and variance. - May cause bias in the estimates.
5	Mode	- It is easy to implement - Suitable for categorical data. - Appropriate for a small number of missing values.	- It favors most frequent value. - May cause changes in the co-variance and variance. - May cause bias in the estimates.
6	LOCF	- It is easy to understand and implement. - Appropriate for time-series data. - Appropriate for a small number of missing values.	- May cause bias in the estimates. - It underestimates the variability of the results.
7	MI	- It restores the natural variability of the missing values. - It incorporates the uncertainty nature of the missing data. - It produces a valid statistical inference. - It produces unbiased estimates	- It is difficult to implement. - It requires large sample size to produce stable estimates - It gives slightly different estimates each time.

		of all parameters. - It is robust. - It is suitable for small sample size or a high number of missing data.	
8	Regression	- It maintains the sample size of the data. - It reduces the standard error.	- It requires large sample size to produce stable estimates.
9	Hot-Deck	- It better approximates the standard deviation of the imputed values. - It preserves the population distribution.	- May not work when there exists no correlation among the features. - It is difficult to implement. - It requires large sample size to produce stable estimates.
10	KNN	- Smaller bias for linear relationship between the variables. - It preserves the population distribution. - It can handle any types of missing data. - It is easy to implement	- It could be computationally expensive with more variables. - It requires large sample size to produce stable estimates.
11	Maximum Likelihood	- It produces unbiased parameter estimates and standard errors. - It is robust. - It is suitable for a high number of missing data.	- It is difficult to implement. - The covariance matrix may be indefinite. - It requires large sample size to produce stable estimates.
12	EM	- It is guaranteed to converge to a local maximum. - It generally outperforms popular single imputation methods. - It produces unbiased parameter estimates and standard errors. - It is robust. - It is suitable for a high number of missing data.	- Might take a long time to converge. - It is difficult to implement. - It requires large sample size to produce stable estimates. - It is model specific.
13	ANN	- It is suitable for a high number of missing data. - It is adaptive to interactions and	- It is difficult to implement. - It requires large sample size to produce stable estimates.

		nonlinearity. - It produces unbiased parameter estimates and standard errors. - It can handle any types of missing data. - It is robust. - It needs little domain knowledge to impute the missing data. - It is guaranteed to perform well in most problems.	
14	DL	- It is suitable for a high number of missing data. - It is adaptive to interactions and nonlinearity. - It produces unbiased parameter estimates and standard errors. - It can handle any types of missing data. - It is robust. - It needs little domain knowledge to impute the missing data. - It is guaranteed to perform well in most problems.	- It is difficult to implement. - It requires large sample size to produce stable estimates.
15	GA	- It is suitable for a high number of missing data. - It is adaptive to interactions and nonlinearity. - It produces unbiased parameter estimates and standard errors. - It is robust. - It is guaranteed to perform well in most problems. - It can handle any types of missing data.	- It is difficult to implement. - It requires large sample size to produce stable estimates.
16	PSO	- It is suitable for a high number of missing data. - It is adaptive to interactions and nonlinearity. - It produces unbiased parameter estimates and standard errors.	- It is difficult to implement. - It requires large sample size to produce stable estimates.

		- It is robust. - It is guaranteed to perform well in most problems. - It can handle any types of missing data.	
17	Matrix Completion	- It is suitable for a high number of missing data. - It is adaptive to interactions and nonlinearity. - It produces unbiased parameter estimates and standard errors.	- It is difficult to implement. - It requires large sample size to produce stable estimates.
18	Decision Tree	- It can handle any types of missing data. - It is adaptive to interactions and nonlinearity. - It is suitable for a high number of missing data. - It produces unbiased parameter estimates and standard errors. - It is guaranteed to perform well in most problems.	- It is difficult to implement. - It requires large sample size to produce stable estimates

Fourth, some practical steps in deciding which missing data handling techniques to use are as follows: (i) Check the percentages of the missing values in each dataset variable and determine the level of importance of the variables containing the missing values. This can be done by looking into the correlation between the independent and target variables. (ii) If a variable(s) is not important to the analysis and it contains some missing values, then it can be deleted (that is, pairwise deletion). However, if a variable(s) is important and contains missing values, then the consideration on which imputation techniques to use is needed. (iii) Select at least any three imputation techniques of interest to test on the dataset and the one which best satisfies the following criteria - unbiased estimates, retains the sample size and reduces standard errors - should be taken. While some of the imputation techniques mentioned above readily meet these criteria, others could also meet these criteria in specific situations.

Sixth, recent research is coming up with some state-of-art methods, such as artificial intelligence (like neural networks) or optimization algorithms (like GA and PSO) or the combination of both, for the approximation of missing data [81, 123-125]. These methods are increasingly being used to handle missing data because they perform appreciably well in highly non-linear scenarios.

Seventh, in Table 4, a summary of the missing data handling techniques, mentioned in Section 3, based on their suitability of the data types and missing data mechanism they can handle, is given.

Table 4: A Summary of the Missing Data Handling Techniques

S/N	Imputation Method	Suitable Data Types	Suitable Missing Mechanism
1	Listwise Deletion	Numerical/Categorical	MCAR
2	Pairwise Deletion	Numerical/Categorical	MCAR
3	Mean	Numerical	MAR
4	Median	Numerical	MAR
5	Mode	Numerical/Categorical	MAR
6	LOCF	Numerical (Time series)	MAR
7	MI	Numerical	MAR, NMAR
8	Regression	Numerical	MAR
9	Hot-Deck	Numerical	MAR
10	KNN	Numerical	MAR
11	Maximum Likelihood	Numerical	MAR, NMAR
12	EM	Numerical	MAR, NMAR
13	ANN	Numerical/Categorical	MAR, NMAR
14	DL	Numerical/Categorical	MAR, NMAR
15	GA	Numerical/Categorical	MAR, NMAR
16	PSO	Numerical/Categorical	MAR, NMAR
17	Matrix Completion	Numerical/Categorical	MAR, NMAR
18	Decision Tree	Numerical/Categorical	MAR, NMAR

7. CONCLUSION

This study reviews some techniques employed in handling missing data in Machine learning projects. It has been established that missing data handling techniques, specifically the imputation methods, are useful and relevant in all fields – engineering, healthcare, e-commerce, finance etc. It is good to note that there is no single best way to handle missing data. Experimentation of different techniques is necessary to decide which technique is best for a particular missing data problem based on the evaluation metrics employed.

However, since each of these missing data handling techniques discussed above have assumed some statistical properties or the other, it should be noted that some model-based, artificial intelligence or optimization algorithm imputations are less biased and produced less errors than most other methods. The lack of use of these high performing imputation methods may be due to poor familiarity and some misconceptions of researchers about them. It could also be due to their computational complexity and lack of programming tools or packages to implement the algorithms compared to other missing data handling techniques.

ACKNOWLEDGMENT

The authors would like to thank the University of Johannesburg, South Africa for supporting this research project.

CONFLICT OF INTERESTS

The authors would like to confirm that there is no conflict of interests associated with this publication and there is no financial fund for this work that can affect the research outcomes.

REFERENCES

[1] Kang H. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 2013; 64(5); 402-406.

[2] Yan T. and Curtin R. The relation between unit nonresponse and item non-response: A response continuum perspective. *International Journal of Public Opinion Research*, 2010; 22(4); 535-551.

[3] De Leeuw E.D., Hox J.J. and Huisman M. Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 2003; 19; 153-176.

[4] Chaitanya Baweja. Employees.csv. Available at: https://github.com/ChaitanyaBaweja/Programming-Tutorials/blob/master/Missing-Data-Pandas/employees.csv, Accessed on 21 December 2020.

[5] Ochieng'Odhiambo F. Comparative study of various methods of handling missing data. *Mathematical Modelling and Applications*, 2020; 5(2); 87-93.

[6] Little R.J.A. and Rubin D.B. (2002) Statistical Analysis with Missing Data. John Wiley & Sons, Inc., USA.

[7] Song Q. and Shepperd M. Missing data imputation techniques. *International journal of business intelligence and data mining*, 2(3); 261-291; 2007.

[8] Makaba T. and Dogo E. A comparison of strategies for missing values in data on machine learning classification algorithms. *In 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2019, pp. 1-7.

[9] Sim J., Lee J.S. and Kwon O. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Mathematical problems in engineering*, 2015; 2015; 1-14.

[10] Stef Van Buuren. (2018) Flexible imputation of missing data. CRC press, USA.

[11] IBM Support. (2020) Pairwise vs. listwise deletion: What are they and when should I use them? Available at: https://www.ibm.com/support/pages/pairwise-vs-listwise-deletion-what-are-they-and-when-should-i-use-them Accessed on 21 December 2020.

[12] Houari R., Bounceur A., Tari A.K. and Kecha M.T. Handling missing data problems with sampling methods. *In 2014 International Conference on Advanced Networking Distributed Systems and Applications*, 2014, pp. 99-104.

[13] Hadeed S.J, O'Rourke M.K., Burgess J.L, Harris R.B. and Canales R.A. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, 2020; 730; 139140.

[14] Malarvizhi M.R. and Thanamani A.S. Comparison of imputation techniques after classifying the dataset using KNN classifier for the imputation of missing data. *International Journal of Computational Engineering Research*, 2013; 3(1); 101-104.

[15] Zhang Z. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 2016; 4(1); 1-8.

[16] Lachin J.M. Fallacies of last observation carried forward analyses. *Clinical trials*, 2016; 13(2); 161-168.

[17] Kenward M.G. and Molenberghs G. Last observation carried forward: a crystal ball? *Journal of biopharmaceutical statistics*, 2009; 19(5); 872-888.

[18] Harel O. and Zhou X. Multiple imputation: review of theory, implementation, and software. *Statistics in medicine*, 2007; 26(16); 3057-3077.

[19] Rubin D.B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 1996; 91(434); 473-489.

[20] Schafer J. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 2002; 57; 19–35.

[21] Schafer J.L. (1997) Analysis of incomplete multivariate data. CRC press, USA.

[22] Rubin D.B. (2004) Multiple imputation for nonresponse in surveys. John Wiley & Sons, USA.

[23] Royston P. and White I.R. Multiple imputation by chained equations (mice): implementation in stata. *J Stat Softw*, 2011; 45(4); 1-20.

[24] Schafer J. L. Multiple imputation: a primer. *Statistical methods in medical research*, 1999; 8(1); 3-15.

[25] Honaker J., King G. and Blackwell M. Amelia II: A program for missing data. *Journal of statistical software*, 2011; 45(7); 1-47.

[26] Raghunathan T., Solenberger P., Berglund P. and Hoewyk J.V. (2016) Iveware: Imputation and variance estimation software (version 0.3). *University of Michigan*.

[27] Graham J. W. Multiple imputation and analysis with SPSS 17-20. (2012) In Missing Data. Springer, Germany.

[28] Ginkel J.R.V. and Ark L.A.V.D. Spss syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, 2005; 29(2); 152-153.

[29] Yuan Y. Multiple imputation using SAS software. *J Stat Softw*, 2011; 45(6); 1-25.

[30] Azur M.J., Stuart E.A., Frangakis C. and Leaf P.J. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 2011; 20(1); 40-49.

[31] Wulff J.N. and Ejlskov L. Multiple imputation by chained equations in praxis: Guidelines and review. *Electronic Journal of Business Research Methods*, 2017; 15(1); 41-55.

[32] Buuren S. V. and Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 2010; 45(3); 1-68.

[33] White I.R., Royston P. and Wood A.M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 2011; 30(4); 377-399.

[34] Bouhlila D.S. and Sellaouti F. Multiple imputation using chained equations for missing data in timss: a case study. *Large-scale Assessments in Education*, 2013; 1(1); 1-33.

[35] Liu X., Ma X., Meng X., Li X. and Xie G. IEEE ICHI data analytics challenge on missing data imputation by amelia II. *In 2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 2019, pp. 1-2.

[36] Pampaka M., Hutcheson G. and Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*, 2016; 39(1); 19-37.

[37] Buuren S.V., Brand J.P.L., Groothuis-Oudshoorn C.G.M. and Rubin D.B. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 2006; 76(12); 1049-1064.

[38] Lee K.J. and Carlin J.B. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology*, 2010; 171(5); 624-632.

[39] Song J. and Belin T.R. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in medicine*, 2004; 23(18); 2827-2843.

[40] Rubin D.B. and Schafer J.L. Efficiently creating multiple imputations for incomplete multivariate normal data. *In Proceedings of the Statistical Computing Section of the American Statistical Association*, 1990, pp. 83-88.

[41] Liu Y. and De A. Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International journal of statistics in medical research*, 2015; 4(3); 287-295.

[42] Schafer J.L. and Graham J.W. Missing data: our view of the state of the art. *Psychological methods*, 2002; 7(2); 147-177.

[43] Horton N.J. and Kleinman K.P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 2007; 61(1); 79-90.

[44] Xu X., Xia L., Zhang Q., Wu S., Wu M. and Liu H. The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology*, 2020; 20(1); 1-9.

[45] Lokupitiya R.S., Lokupitiya E. and Paustian K. Comparison of missing value imputation methods for crop yield data. Environmetrics: *The official journal of the International Environmetrics Society*, 2006; 17(4); 339-349.

[46] Ford B.L. (1983) An overview of hot-deck procedures. Incomplete data in sample surveys, Academic Press, USA.

[47] Andridge R.R. and Little R.J.A. A review of hot deck imputation for survey non-response. *International statistical review*, 2010; 78(1); 40-64.

[48] Liao S.G., Lin Y., Kang D.D., Chandra D., Bon J., Kaminski N., Sciurba F.C. and Tseng G.C. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC bioinformatics*, 2014; 15(1); 1-12.

[49] Ding Y. and Ross A. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, 2012; 45(3); 919-933.

[50] Deza M.M. and Deza E. (2006) Dictionary of distances. Elsevier, Netherlands.

[51] Kim K.Y., Kim B.J. and Yi G.S. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 2004; 5(1); 1-9.

[52] Caruana R. A non-parametric em-style algorithm for imputing missing values. *In International Workshop on Artificial Intelligence and Statistics*, 2001, pp. 35-40.

[53] Brás L.P. and Menezes J.C. Improving cluster-based missing value estimation of dna microarray data. *Biomolecular engineering*, 2007; 24(2); 273-282.

[54] Kim H., Golub G.H and Park H. Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 2006; 22(11); 1410-1411.

[55] Zhang X., Song X., Wang H. and Zhang H. Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 2008; 38(10); 1112-1120.

[56] Cai Z., Heydari M. and Lin G. Iteratedc local least squares microarray missing value imputation. *Journal of bioinformatics and computational biology*, 2006; 4(5); 935-957.

[57] Jadhav A., Pramod D. and Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 2019; 33(10); 913-933.

[58] Zainuri N.A., Jemain A.A. and Muda N. A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana*, 2015; 44(3); 449-456.

[59] Dongare A.D., Kharde R.R. and Kachare A.D. Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2012; 2(1); 189-194.

[60] Mishra M. and Srivastava M. A view of artificial neural network. *In 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014)*, 2014, pp. 1-3.

[61] Abiodun O.I., Jantan A., Omolara A.E., Dada K.V., Mohamed N.A. and Arshad H. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 2018; 4(11); e00938.

[62] Smieja M., Struski L., Tabor J., Zieliński B. and Spurek P. Processing of missing data by neural networks. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018, pp. 1-11.

[63] Choudhury S.J. and Pal N.R. Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, 2019; 182; 104838.

[64] Goodfellow I., Bengio Y. and Courville A. (2016) Deep Learning. MIT Press, USA.

[65] Raaijmakers S. Deep Learning for Natural Language Processing. (2022) Leiden University, Netherlands.

[66] Cheng C.Y., Tseng W.L., Chang C.F., Chang C.H. and Gau S.S.F. A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder. *Frontiers in psychiatry*, 2020; 11; 673.

[67] Duan Y., Lv Y., Kang W. and Zhao Y. A deep learning based approach for traffic data imputation. *In 17th International IEEE conference on intelligent transportation systems (ITSC)*, 2014, pp. 912-917.

[68] Yoon S. and Sull S. Gamin: generative adversarial multiple imputation network for highly missing data. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8456-8464.

[69] Kim J., Tae D. and Seok J. A survey of missing data imputation using generative adversarial networks. *In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 454-456.

[70] Qiu Y.L., Zheng H. and Gevaert O. A deep learning framework for imputing missing values in genomic data. BioRxiv, 2018, p. 406066.

[71] Beaulieu-Jones B.K. and Jason H Moore J.H. Missing data imputation in the electronic health record using deeply learned autoencoders. *In Pacific symposium on biocomputing*, 2017, pp. 207-218.

[72] Yoon J., James J. and Schaar M.V.D. Missing data imputation using generative adversarial nets. *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5689-5698.

[73] Li S.C.X., Jiang B. and Marlin B. Misgan: Learning from incomplete data with generative adversarial networks. *Seventh International Conference on Learning Representations (ICLR 2019)*, 2019, pp. 1-20.

[74] Katoch S., Chauhan S.S. and Kumar V. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 2021; 80(5); 8091-8126.

[75] Sulejmani A. and Koça O. Development of Optimal Transmission Rate of the Kinematic Chain by using Genetic Algorithms Coded in Mathcad. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 2021; 4(4); 792-803.

[76] Holland J.H. (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, USA.

[77] Sampson J.R. (1976) Adaptation in natural and artificial systems University of Michigan Press, USA.

[78] Haldurai L., Madhubala T. and Rajalakshmi R. A study on genetic algorithm and its applications. *International Journal of computer sciences and Engineering*, 2016; 4(10); 139-143.

[79] Maulik U. and Bandyopadhyay S. Genetic algorithm-based clustering technique. *Pattern recognition*, 2000; 33(9); 1455-1465.

[80] Shahzad W., Rehman Q. and Ahmed E. Missing data imputation using genetic algorithm for supervised learning. *International Journal of Advanced Computer Science and Applications*, 2017; 8(3); 438-445.

[81] Noei M. and Abadeh M.S. A genetic asexual reproduction optimization algorithm for imputing missing values. *In 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2019, pp. 214-218.

[82] Priya R.D., Kuppuswami S. and Kumar S.M. A genetic algorithm approach for non-ignorable missing data. *International Journal of Computer Applications*, 2011; 20(4); 37-41.

[83] Priya R.D. and Sivaraj R. Dynamic genetic algorithm-based feature selection and incomplete value imputation for microarray classification. *Current Science*, 2017; 112(1); 126-131.

[84] Elzeki O.M., Alrahmawy M.F. and Elmougy S. A new hybrid genetic and information gain algorithm for imputing missing values in cancer genes datasets. *International Journal of Intelligent Systems and Applications*, 2019; 11(12); 20-33.

[85] Lobato F., Sales C., Araujo I., Tadaiesky V., Dias L., Ramos L. and Santana A. Multi-objective genetic algorithm for missing data imputation. *Pattern Recognition Letters*, 2015; 68; 126-131.

[86] Eberhart R. and Kennedy J. A new optimizer using particle swarm theory. In MHS'95. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39–43.

[87] Wang D., Tan D. and Liu L. Particle swarm optimization algorithm: an overview. *Soft Computing*, 2018; 22(2); 387-408.

[88] Krishna M. and Ravi V. Particle swarm optimization and covariance matrix based data imputation. *In 2013 IEEE International Conference on Computational Intelligence and Computing Research*, 2013, pp. 1-6.

[89] Gautam C. and Ravi V. Data imputation via evolutionary computation, clustering, and a neural network. *Neurocomputing*, 2015; 156; 134-142.

[90] Salleh M.N.M. and Samat N.A. Fcmpso: An imputation for missing data features in heart disease classification. *In IOP Conference Series: Materials Science and Engineering*, 2017; 226; 012102.

[91] Candès E.J. and Recht B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009; 9(6); 717-772.

[92] Johnson C.R. Matrix completion problems: a survey. *In Matrix theory and applications*, 1990; 40; 171-198.

[93] Candès E.J. and Tao T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 2010; 56(5); 2053-2080.

[94] Genes C. (2018) Novel Matrix Completion Methods for Missing Data Recovery in Urban Systems. PhD thesis, University of Sheffield.

[95] Cai J.F., Candès E.J. and Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 2010; 20(4); 1956-1982.

[96] Chatterjee S. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 2015; 43(1); 177-214.

[97] Schnabel T., Swaminathan A., Singh A., Chandak N. and Joachims T. Recommendations as treatments: Debiasing learning and evaluation. *In international conference on machine learning*, 2016, pp. 1670-1679.

[98] Zheng X., Wang M., Xu R., Li J. and Wang Y. Modeling dynamic missingness of implicit feedback for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2020; 34(1); 405-418.

[99] Ma W. and Chen G.H. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019, pp. 1-10.

[100] Gavankar S. and 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)*, 2015, pp. 122-126.

[101] Twala B. and Cartwright M. Ensemble missing data techniques for software effort prediction. *Intelligent Data Analysis*, 2010; 14(3); 299-331.

[102] Twala B., Jones M.C. and Hand D.J. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 2008; 29(7); 950-956.

[103] Song Y.Y. and Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 2015; 27(2); 130-135.

[104] Durmuş B. and Güneri Ö. İ. Investigation of Factors Affecting Immunotherapy Treatment Results by Binary Logistic Regression and Classification Analysis. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 2020; 3(3), 467-473.

[105] Tang F. and Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: *The ASA Data Science Journal*, 2017; 10(6); 363-377.

[106] Davydenko A. and Fildes R. (2016) Forecast error measures: critical review and practical recommendations. Business Forecasting: Practical Problems and Solutions. Wiley, USA.

[107] Chen C., Twycross J. and Garibaldi J.M. A new accuracy measure based on bounded relative error for time series forecasting. *PloS one*, 2017; 12(3); e0174202.

[108] Botchkarev A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. *SSRN*, 2018;1-16.

[109] Shcherbakov M.V., Brebels A., Shcherbakova N.L., Tyukov A.P., Janovsky T.A. and Kamaev V.A. A survey of forecast error measures. *World Applied Sciences Journal*, 2013; 24(24); 171-176.

[110] Schmitt P., Mandel J. and Guedj M. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 2015; 6(1); 1-6.

[111] Li Y., Li Z., and Li L. Missing traffic data: comparison of imputation methods. *IET Intelligent Transport Systems*, 2014; 8(1); 51-57.

[112] Nur-E-Arefin M. A Comparative Study of Machine Learning Classifiers for Credit Card Fraud Detection. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 2020; 3(1), 395-406.

[113] Hyndman R.J. and Koehler A.B. Another look at measures of forecast accuracy. *International journal of forecasting*, 2006; 22(4); 679-688.

[114] Botchkarev A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology, *Interdisciplinary Journal of Information, Knowledge, and Management*, 2019; 14; 45-79.

[115] Pascual C. Tutorial: Understanding regression error metrics in python. Available at: https://www.dataquest.io/blog/understanding-regression-error-metrics/, Accessed on 21 December 2021.

[116] Junninen H., Niska H., Tuppurainen K., Ruuskanen J. and Kolehmainen M. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 2004; 38(18); 2895-2907.

[117] Moepya S.O., Akhoury S.S., Nelwamondo F.V. and Twala B. The role of imputation in detecting fraudulent financial reporting. *International Journal of Innovative Computing, Information and Control*, 2016; 12(1); 333-356.

[118] Beauxis-Aussalet E. and Hardman L. Visualization of confusion matrix for non-expert users. *In IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*, 2014, pp. 1-2.

[119] Le T.D., Beuran R. and Tan Y. Comparison of the most influential missing data imputation algorithms for healthcare. *In 2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 247-251.

[120] Cihan P., Kalıpsız O. and Gökçe E. Effect of imputation methods in the classifier performance. Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2019; 23(6); 1225-1236.

[121] Xu X., Chong W., Li S., Arabo A. and Xiao J. Miaec: Missing data imputation based on the evidence chain. *IEEE Access*, 2018; 6:12983-12992.

[122] Sefidian A.M. and Daneshpour N. Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, 2019; 115; 68-94.

[123] Abdella M. and Marwala T. Treatment of missing data using neural networks and genetic algorithms. *In Proceedings. IEEE International Joint Conference on Neural Networks*, 2005; 1; 598-603.

[124] Leke C. and Marwala T. Missing data estimation in high-dimensional datasets: A swarm intelligence-deep neural network approach. *In International Conference on Swarm Intelligence*, 2016, pp. 259-270.

[125] Leke C., Twala B. and Marwala T. Modeling of missing data prediction: Computational intelligence and optimization algorithms. *In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 1400-1404.