Accepted Manuscript

Vietnam Journal of Computer Science

Article Title:	The effects of missing data characteristics on the choice of imputation techniques
Author(s):	Oyekale Abel Alade, Ali Selamat, Roselina Salehuddin
DOI:	10.1142/S2196888820500098
Received:	15 October 2018
Accepted:	17 January 2020
To be cited as:	Oyekale Abel Alade, Ali Selamat, Roselina Salehuddin, The effects of missing data characteristics on the choice of imputation techniques, *Vietnam Journal of Computer Science*, doi: 10.1142/S2196888820500098
Link to final version:	https://doi.org/10.1142/S2196888820500098

International Journal of Information Acquisition

THE EFFECTS OF MISSING DATA CHARACTERISTICS ON THE CHOICE OF IMPUTATION TECHNIQUES

Oyekale Abel Alade
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia
kaleabel@gmail.com

Ali Selamat*
Media & Games Center of Excellence (MAGICX) & School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia
Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic.
aselamat@utm.my

Roselina Sallehuddin
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia
roselina@utm.my

One major characteristic of data is completeness. Missing data is a significant problem in medical datasets. It leads to incorrect classification of patients and dangerous to the health management of patients. Many factors lead to the missingness of values in databases in medical datasets. In this paper, we propose the need to examine the causes of missing data in a medical dataset to ensure that the right imputation method is used in solving the problem. The mechanism of missingness in datasets was studied to know the missing pattern of datasets and determine the suitable imputation technique to generate complete datasets. The pattern shows that the missingness of the dataset used in this study is not a monotone missing pattern. Also, single imputation techniques underestimate variance and ignore relationships among variables; therefore, we used multiple imputation techniques that run in 5 iterations for the imputation of each missing value. The whole missing values in the dataset were regenerated 100%. The imputed datasets were validated using an extreme learning (ELM) classifier. The results show improvement in the accuracy of the imputed datasets. The work can, however, be extended to compare the accuracy of the imputed datasets with the original dataset with different classifiers like support vector machine (SVM), radial basis function (RBF), and extreme learning machines (ELM).

Keywords: Imputation techniques, mechanism of missingness, missing data, missing pattern, multiple imputations.

1. Introduction

Missing data/values describe the absence of important data items in instances of datasets. Data are collected at various points for medical investigation. Lichman [2013] observed two possible types of databases in the medical domain.

*Correspondence author. Ali Selamat is currently a visiting professor at the University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic, EU. Email: aselamat@utm.my

The first type is basically for hospital information systems, which consists of a vast number of attributes. Many of such attributes are not directly required for the diagnosis of the ailments in the patients. The other type of medical database is collected by experts. The databases may contain unique research data on topics based on hypothesis propositions that must be investigated. Missing data is very much pervasive in either of these types of databases, as in many other databases, and most real-world data analysis tasks. García-Laencina, Sancho-Gómez, & Figueiras-Vidal, [2010]; Tran, Zhang, Andreae, Xue, & Bui [2017] observed that 45% of datasets in UCI machine learning repository are marred by missing values, most of which may fall in the category of medical data. The pervasive nature of missing data has been described as one of the most challenging tasks in data science [Baraldi & Enders, 2010]. The missing observations usually have the potential to be captured but were not captured due to some reasons that may arise from the patient or the medical personnel. In Gao, Liu, Peng, & Jian [2015], two basic patterns of missing data were considered. The patterns are: (a) missing features, a situation when the features exist but values were not taken; therefore the information is lost, or the cost of acquiring the feature is high; and (b) missing label, when the label is inherently missing – that is, a problem that cannot be avoided - which eventually affects the performance of classifiers in the face of ever-growing datasets [Gimpy, 2014] in this information age. Missing data affects some operations in medical research. It makes it challenging to extract useful information from datasets; also, feature selection is not always applied to datasets with missing values [Tran *et al*., 2017].

Missing data is a complex pattern problem that is inherent in equipment malfunction during data extractions, sampling, transcription, transmission, and noise during data pre-processing [Gao *et al*., 2015]. Newgard & Lewis [2015] observed that missingness of data in clinical research could be a result of variables that are complex, time-sensitive, resource-intensive, or method of collection of longitudinal data. There are other causes of missing values in datasets observed in [Gimpy, 2014]. They are (i) Possibility of the observations been irrelevant, especially in medical data collection, because the primary reason for data collection is for medical investigation and diagnosis, and not really for research purposes. (ii) Inability to record the values when they were collected due to the emergency situation; or patient response avoidance from the respondent for privacy. (iii) The omission of essential features during the data collection plan. (iv) Non-capturing of seemingly available values. Two significant problems envisaged in Zhu [2014] with the presence of a missing value in datasets are (i) reduction in overall statistical power and (ii) statistical bias estimation.

The problem of missing data is being debated for some time [Joseph, 2016]. The best method to fix missing values is to revisit the data collection/extraction process to recollect and correct missing values and noisy values, respectively. This may involve the re-investigation of patients from various examination units to fix the missing values, but this method of recollection/re-extraction might not be practicable. Therefore, there is a need for fitting techniques with close approximations to the real data. The resolve, however, is that the imputation of missing data has no definite solution [Kenward, 2013].

The accuracy of classification from the incomplete dataset is unreliable because some vital information relevant to the analysis might be lost. Besides, some classifiers find it challenging to run in the face of missing values [Gautam & Ravi, 2015]. Therefore, there is a necessity for missing value imputation. The focus of imputation is to estimate the possible value of the missing data from observed data and fixed the estimated values as a replacement of the lost values. That is, to ensure complete datasets.

Several approaches have been used in research for the treatments of missing data in datasets. Some authors used the percentage of missingness in a dataset as a measure for the choice of imputation technique [Joseph, 2016]. Some implementations treat missing data implicitly. This brings about

different results when such treatments are replicated using different applications. Although the difference may not be significant, however, these approaches compromise the scientific soundness of the studies. An explicit approach to handling missing data is a better practice. In Zhu [2014], it was observed that the choice of missing data handling imputation technique depends on the research focus, whether it is a pragmatic or an analytical approach. Although missing data is pervasive in data science generally, in this paper, we focus the study on the effect of missing values on the medical datasets as an offshoot to application areas.

From this point on, the paper is organized as follows: Section 2 reviews relevant work on medical datasets with missing data and some techniques of missing data imputation; section 3 describes the characteristics of missing values, that is, mechanism of missingness; various treatments of missing values in datasets are discussed in sections 4; section 5 explains the proposed multiple imputation model; section 6 reports the experimental setup; discussion of the results of imputation on Pima Indian diabetes dataset is in 7; while section 8 concludes the paper.

2. Review of Literature

Medical datasets have been classified by many researchers. Some of the datasets are complete, while some are incomplete. Several studies had been carried out on missing data and imputation techniques from different perspectives. Some authors explicitly treat missing values using different imputation techniques, while some are passive about it, leading to the assignment of zeros (0), deletions of cases/features, or completely ignore missing values. ELM is widely used in recent time to solving classification, clustering, compression, forecasting, and regression problems [Alade, Selamat, & Sallehuddin, 2018] because it tolerates quite a good number of feature mapping functions such as sigmoid, hard-limit, Gaussian, multi-quadratic, wavelet, Fourier series, etc., and it handles large and small datasets efficiently [Huang, 2015]. Subbulakshmi & Deepa [2015] proposed a machine learning paradigm by integrating particle swarm optimization (PSO) technique with extreme learning machines (ELM) to classify some medical datasets. The hybrid system performs well compare to other classifiers; however, missing values in the datasets were substituted with zeros. This approach is scientifically unfit for accurate results. Zeros do not represent a good imputation of missing values.

Bai, Mangathayaru, & Rani [2015] overviewed the hidden challenges of missing values in medical datasets during pre-processing. They proposed the imputation of the missing values in the medical datasets with categorical attributes, the causes and pattern of missingness in the datasets were, however, not considered. This may result in the wrong choice of imputation technique.

An extensive review was carried out in [Armina, Mohd Zain, Ali, & Sallehuddin, 2017] on missing value imputations. The authors provide a detailed analysis of various imputation techniques. They grouped imputation techniques into four (4) broad categories: global, local, hybrid, and knowledge assisted approaches, but there was no experiment conducted to prove any of the imputation techniques discussed in their study. Gaussian mixture model and extreme learning machines (GMM-ELM) was proposed as a reliable approximation technique for imputing missing data by Sovilj *et al.* in [Sovilj *et al*., 2015]. The results of their work improved the imputation of missing values over the mean imputation technique; however, the execution time was longer. Bai *et al*. [2015] addressed categorical attribute missing values in medical datasets using imputation measure. The work, however, used a hypothetical dataset with only nine (9) cases and only two (2) missing values; the characteristics of the missingness was not considered in work. In Tsai & Chang [2016], Tsai & Chang investigated the effects of filtering outliers from datasets on imputation tasks using instance selection on categorical, numerical, and mixed type attributes. The effectiveness of the method was tested with k-NN and SVM classifiers. To compare the performance of three Bayesian imputation techniques, [Austin & Escobar, 2005] placed prior distribution on attributes with missing values. Monte

Carlo simulation model was used to examine the performance of the sibling models. The result showed that mean and mean square error of logistic and Bayesian models depend on risk factors examined, and the mechanism of missing data been used. This gives an insight into the necessity for consideration of the mechanism of missingness for the right choice of imputation technique. Multiple imputations with Pohar-Perme method was used in [Falcaro & Carpenter, 2017] to estimate the net survival for stage-specific colorectal cancer. They concluded that the interpretation of datasets with a high percentage of missing values should be cautious and should be with sensitivity analysis. However, the characteristics of the missingness of data in the dataset were not taken into consideration before the choice of the imputation technique.

In Tang, Zhang, Wang, Wang, & Liu [2015], a hybrid imputation method based on the integration of Fuzzy C-Means (FCM) and the Genetic Algorithm (GA) for missing traffic volume data was developed. The study based the estimation on inductance loop detector outputs. The result, under prevailing traffic conditions, performed better than conventional methods. All these methods and much more in literature underscore the need for imputation of missing data in a given dataset with missing values.

Although most of the imputation techniques mentioned above attempt to fill the missing values by approximately conforming to the distribution of the datasets, however, the methods of the imputation of values are not explicitly modeled; therefore, further analysis is ignored [Sovilj *et al*., 2015] and thereby lead to bias result. Nguyen, Carlin, & Lee [2017] raised some critical points to consider when constructing an imputation model. These are (a) model imputation functional form, (b) feature selection for the model, (c) inclusion of non-linear relationships in the model, and (d) the best way to handle non-normal continuous features. Nguyen concluded that there is no consensus in literature on how to implement these decisions, these could be evaluated from the nature of the missing values in the datasets. Therefore, there is a need to know the nature of missingness in a dataset for the right choice of imputation technique. In the next section, we attempt to have an overview of the possible nature of missingness in datasets.

3. Mechanisms of Missingness

The focus of this section is to look at the characteristics of missing values in datasets. These characteristics determine the causes of the missingness in the dataset. It is good to know the cause(s) of missing values in a dataset to handle the missingness appropriately [Liu & Gopalakrishnan, 2017]. Some literature refers to this as *mechanisms of missingness*. Various mechanisms of missing data values abound in literature [Diaconis & Efron, 1983; Falcaro & Carpenter, 2017; Huang & Chen, 2007; Shang & He, 2015; Zhu, 2014], but the most popular ones are basically three (3) which shall be considered for the purpose of this study.

3.1. *Missing at Random (MAR)*

This is a type of missingness that does not occur entirely at random; instead, they occur where there are other variables with complete information that can account for the missingness. It does not necessarily mean that the cases are similar to the complete counterpart. MAR is more realistic than missing completely at random (MCAR), and it is mostly applied to missing data imputation in many pieces of literature [Newgard & Lewis, 2015]. It is based on an ignorable assumption: that is, the available information is sufficient, and the assignment mechanism can be ignored. This case arises when some respondents decide to hide some information that is personal or are unpopular about themselves [Wasito & Mirkin, 2006]. Logistic regression with the outcome of 1 for the observed and 0 for the missing values is a reasonable option for its treatment. It can be statistically expressed as in (1) thus: for X random attribute and Z predictor attribute, if

$$P(X|x_{miss}) = P(X|x_{obs}, Z) \quad (1)$$

then x distribution is not affected by values

$$X \in Z$$

That is, when the missingness is based on the observed factors, then it is independent of the unobserved factors.

Although, MAR in popularly accepted in many techniques, the result is still bias or yield imprecise results with simple imputation techniques [Newgard & Lewis, 2015].

3.2. *Missing Completely at Random (MCAR)*

MACR occurs if the cause of missing values of observable features and the parameters of unobservable features of interest are independent, and their occurrence is entirely at random. Analyses preform on MCAR datasets are unbiased, although this type of datasets is rare. It is the highest level of randomness. It is expressed as in (2)

$$P(X|x_{miss}) = P(X|x_{obs}) \quad (2)$$

Any imputation technique can be applied [Tsai & Chang, 2016]

3.3. *Missing Not at Random (MNAR)*

MNAR is a type of missing data where there is a relationship between the missing data and the reason for the missingness. It occurs when the missingness depends on the probability of the actual value of the missing data [Gimpy, 2014] and some/all other observed data [Zhu, 2014] [Austin & Escobar, 2005]. Tsai & Chang [2016] observed that this mechanism would be difficult to judge because the missing data are unknown.

The treatments of missing data should be based on the mechanism of missing data, as explained in the next section.

4. Treatments of Missingness

In the treatment of missing data, two broad approaches are conventional in literature. These are (a) omission of missing data, (b) imputing the missing data [Sovilj *et al*., 2015]. Some approaches to the treatment of missing values (MV) are outlined in Fig. 1 and later discussed below:

4.1. *Omission*

This approach simply deletes instances with missing data. The approach is common in some regression models, usually refers to as *Lit-wise deletion, complete case analysis* [Mukaka *et al*., 2016]. It is only valid under the following conditions: (a) the instances with missing values in the sample are negligible; (b) the pattern of the missing data is missing at random (MAR) or missing completely at random [Gautam & Ravi 2015]. This approach reduces the sample size, so they often lost vital information because the deleted instances may be essential and deciding factors for predictions and classifications [Sovilj *et al.,* 2015]. It limits the study power. Therefore, the imputation technique is preferred to lit-wise deletion.

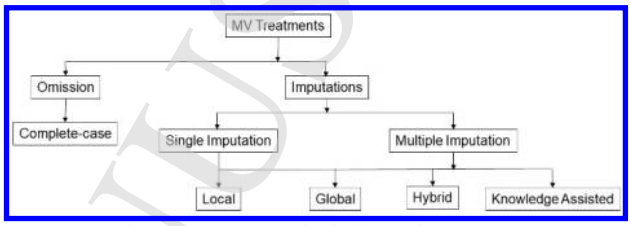

Fig. 1: Treatment of missing values (MV)

4.2. *Imputation Techniques*

Imputation can be categorized into two: (a) single imputation and (b) multiple imputations

4.2.1. *Single Imputation*

Single imputation uses mean, median, mode, or conditional mean like a predicted value from a regression function evaluation or decision tree [Tran *et al*., 2017; Wasito & Mirkin, 2006] to generate the data to be imputed only ones. Mean imputation is a standard method used to replace missing values [Tran *et al*., 2017]. The average or median of the observed feature values is computed and substituted for each missing value for numerical attributes; and mode for simple ones. Unfortunately, this method does not present an actual distribution of features. It

underestimates the variance and ignores relationships among variables in the dataset [Austin, Escobar, 2005], [Falcaro & Carpenter, 2017]. These lead to complications of statistical inferences. Imputation of missing data is handled sequentially – one-by-one. This method work with datasets with a limited number of variables [Wasito & Mirkin, 2006]

The last observation carried forward: this method replaces every missing value with the last observation. This approach assumes that the result will not change after the last reading [Zhu, 2014]. It is a simple method, so it is accessible. It maintains the actual size of data; however, it might result in bias outcomes. This method is not analytical enough, so there is a need for a more comprehensive imputation technique.

4.2.2. *Multiple Imputation Techniques*

This method is scientifically plausible to replace values with the modeled results of several imputations that analytically represent the missing data. For example, the regression model will reflect the uncertainty of regression coefficients and the sample variables in the model. Multiple imputation techniques analytically create several values to replace the missing data [Gelman & Hill, 2007]. Different models also predict this replacement values. It must be known that the aim of multiple imputations is not to produce the actual missing value; rather, it attempts to generate scientifically valid results to account for the missing values [Zhang, 2016]. According to Rodwell, Lee, Romaniuk, & Carlin [2014], it is possible that the simulated values may not fall within the expected range. The single idea about multiple imputations, however, is to form *N* complete datasets from the observed value analytically. *N* is the number of imputations carried out on the original dataset with missing values, and it produces N different complete datasets.

Armina et al. [2017], further details on local, global, hybrid, and knowledge assisted imputation techniques are discussed, as sketched out in Fig. 1. In the next section, we propose a regression model for multiple imputations used in this study.

5. Proposed Multiple Imputation Model

In this section, a regression model is proposed for multiple imputations of missing data, because it draws values randomly from donor instances to predict values that are closed to the missing values to be predicted; also regression creates inter-data variability using stochastic elements [van Kuijk, Viechtbauer, Peeters, & Smits, 2016]. The results of this data pooling produce correct standard error estimates.

For a dataset with missing data problem of Y = X_{m+1} on m variables X_1, …, X_m instances of a random sample when x are incomplete; the incomplete values can be estimated with regression model in (3):

$$E(Y|X_1,\dots,X_m) = \beta_0 + \sum_{m=1}^{M} \beta_m X_m \tag{3}$$

where $X = (X_1,\dots,X_m)$ is the correlation coefficient, $\beta = (\beta_1,\dots,\beta_m)$ the relative size of the coefficient of regression to one another, and β_0 is the intercept.

To make an inference of the regression coefficient, β_m of the missing values and the standard errors $e_1 \dots e_M$ are obtained for each dataset in M. The mean dataset estimates ($\hat{\beta}$) is given in (4):

$$\widehat{\beta} = \frac{1}{m}\sum_{m=1}^{M} \beta_m \tag{4}$$

The variation within (V) and between (W) the imputation is shown in (5) below:

$$V_\beta = W + \left(1 + \tfrac{1}{m}\right) E \tag{5}$$

where

$$W = \frac{1}{m}\sum_{m=1}^{M} e^2, \quad E = \frac{1}{m-1}\sum_{m=1}^{M} (\hat{\beta}_m - \hat{\beta}$$

6. Experimental Setup

In order to determine the choice of imputation techniques for a dataset, the Pima Indians diabetes dataset from UCI was used. Pima Indians diabetes

dataset contains clinical tests and diagnosis of Pima Indian women of 21 years of age and above with diabetes [Smith, Everhart, Dickson, Knowler, & Johannes, 1988; Strack *et al.*, 2014]. The dataset is made up of integer and real number data types. It has 768 instances – 8 predictive features and a class feature. The features in the dataset are number of time pregnant (V1), a 2hr oral tolerant test for plasma glucose concentration (V2), diastolic blood pressure (V3), triceps skinfold thickness (V4), a 2hr serum insulin (V5), body mass index (V6), diabetes pedigree (V7), and age (V8). The class (V9) is a binary classification dataset with 1 for positive and 0 for negative. The dataset was indicated to have missing values on the web page [Smith, Everhart, Dickson, *et al.*, 1988], but the real dataset did not show such signs. However, a critical examination of the dataset shows that variables like plasma glucose concentration, body mass index, triceps of skinfold thickness, diastolic blood pressure, 2-hour serum insulin cannot be 0 for any instance; therefore, it was assumed that all the data values scored 0s are really missing values and were so treated in this study.

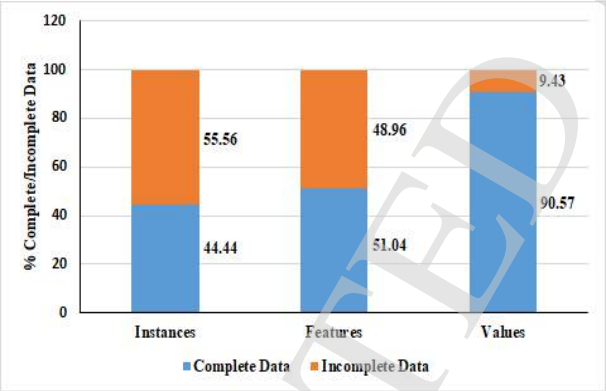

Figure 2: Percentage of complete/incomplete data in instances, features, and values of Pima Indians diabetes dataset

The missing values in the dataset were coded for proper identification by the imputation program. For an analytical presentation of the degree of missingness in the dataset, graphical, and numerical summaries [Nguyen *et al.*, 2017] were used in our report. The percentage of missingness was calculated by instances, features, and values, as shown in Fig. 2.

The percentage of missingness among the instances was considered in this study as common in literature [Austin & Escobar, 2005; Tsai & Chang, 2016]. The missing values were analyzed in order to know the distributions of missingness among the various features that have missing values in the dataset. The distribution is shown in Table 1.

Table 1: Distribution of missing values among the features with incomplete data in the instances of the Pima Indians diabetes dataset.

	Missing N	%	Valid N	Mean	Dev.
V5	74	48.7%	394	155.55	118.776
V4	27	29.6%	541	29.15	10.477
V3	5	4.6%	733	72.41	12.382
V6	1	1.4%	757	32.457	6.9250
V2		0.7%	763	121.69	30.536

A missing pattern describes the set of features in a dataset that has at least an instance in the dataset with a missing value(s) the same feature(s) in the

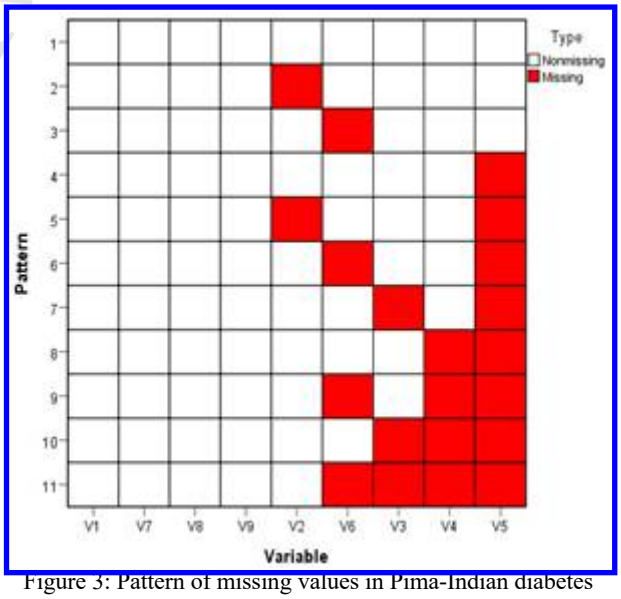

Figure 3: Pattern of missing values in Pima-Indian diabetes

pattern [Tran *et al.*, 2017].

The pattern of missingness of the features in our dataset was analyzed. The result of the analysis is shown in Fig. Each pattern in the figure corresponds

to the group of instances with the same pattern of incomplete and complete data.

The variables are sorted in increasing order of missing values (from the least missing value feature to the one with the highest missing value) from left to right. This is to enable us to know the type of imputation required to fill up the missingness. Multiple imputations are carried out on the original dataset to be able to come up with five sets of complete datasets. The analysis of the imputed datasets for features with missing values is shown in Table 2-6

Table 2*:* Multiple imputations for V2

Imputation	N	Mean	Dev.	Min.	Max.
Original Data	763	121.69	30.536	44.00	199.00
Imputed Values 1	5	125.52	38.578	88.84	170.01
2	5	119.22	10.672	111.21	137.24
3	5	145.21	35.676	104.01	191.63
4	5	111.99	27.614	76.61	140.40
5	5	135.09	27.640	95.38	172.70
Complete Data After Imputation 1	768	121.71	30.565	44.00	199.00
2	768	121.67	30.446	44.00	199.00
3	768	121.84	30.603	44.00	199.00
4	768	121.62	30.511	44.00	199.00
5	768	121.77	30.520	44.00	199.00

Table 3*:* Multiple imputations for V3

Imputation	N	Mean	Dev.	Min.	Max.
Original Data	733	72.41	12.382	24.00	122.00
Imputed Values 1	35	72.55	14.902	43.49	100.49
2	35	72.25	13.165	44.49	97.34
3	35	74.56	12.447	43.87	101.37
4	35	70.85	11.039	45.90	95.96
5	35	69.06	11.674	50.13	96.20
Complete Data After Imputation 1	768	72.41	12.497	24.00	122.00
2	768	72.40	12.410	24.00	122.00
3	768	72.50	12.385	24.00	122.00
4	768	72.33	12.322	24.00	122.00
5	768	72.25	12.363	24.00	122.00

Table 4*:* Multiple imputations for V4

Imputation	N	Mean	Dev.	Min.	Max.
Original Data	541	29.15	10.477	7.00	99.00
Imputed Values 1	227	27.84	10.680	-3.03	63.88
2	227	28.47	10.895	-1.92	55.99
3	227	27.45	10.373	-.62	56.51
4	227	27.85	10.190	-6.94	58.87
5	227	27.74	11.160	.75	57.31
Complete Data After Imputation 1	768	28.77	10.547	-3.03	99.00
2	768	28.95	10.600	-1.92	99.00
3	768	28.65	10.469	-.62	99.00
4	768	28.77	10.403	-6.94	99.00
5	768	28.74	10.696	.75	99.00

Table 5*:* Multiple imputations for V5

Imputation	N	Mean	Dev.	Min.	Max.
Original Data	394	155.55	118.776	14.00	846.00
Imputed Values 1	374	146.07	111.231	-113.11	489.68
2	374	153.85	122.138	-227.40	488.04
3	374	154.91	120.382	-257.32	488.35
4	374	144.23	126.086	-182.66	540.58
5	374	134.68	119.561	-180.91	521.16
Complete Data After Imputation 1	768	150.93	115.186	-113.11	846.00
2	768	154.72	120.349	-227.40	846.00
3	768	155.24	119.483	-257.32	846.00
4	768	150.03	122.441	-182.66	846.00
5	768	145.39	119.538	-180.91	846.00

.

Table 6*:* Multiple imputation for V6

Data	Imputation	N	Mean	Dev.	Min.	Max.
Original Data		757	32.457	6.9250	18.200	67.100
Imputed Values	1	11	32.271	7.4383	18.047	46.451
	2	11	34.026	7.7246	22.682	45.062
	3	11	29.469	7.2017	18.687	43.716
	4	11	32.797	6.2029	22.493	42.757
	5	11	28.430	6.8445	13.355	38.423
Complete Data After Imputation	1	768	32.455	6.9274	18.047	67.100
	2	768	32.480	6.9340	18.200	67.100
	3	768	32.415	6.9333	18.200	67.100
	4	768	32.462	6.9117	18.200	67.100
	5	768	32.400	6.9360	13.355	67.100

7. Discussion of Results

As mentioned earlier, the original Pima Indians diabetes dataset could not be easily noticed as having missing data because every cell in the dataset is completely scored. However, the source (UCI database) categorically stated that the dataset has missing values.

On a critical look, it was observed that the missing values in the dataset were scored zero, and it was so treated except the first feature (Number of times pregnant – V1), which is assumed can be zero among the selected women.

Fig. 2 shows that 44.44% of the features in the data set is complete. That is, 4 out of 9 variables (class label inclusive) have complete data scored while 55.56% have incomplete (missing) values. Based on the instances (cases) in the data set, 51.04%, which is 392 cases, has complete data, while 48.96% (366 cases) have incomplete data. For the entire values in the Pima Indians diabetes dataset (that is the intersections of features and instances values), 90.57% of the cells have values while 9.43% are missing.

In Table 1, the distribution of missing values among the cases is presented. The table shows the five features with missing values in their quantities and in various percentages across the dataset. The features are arranged in descending order of the percentage constituents of their missingness: 2-hour serum insulin (V5) has the highest percentage of missing values (48.7%), triceps skinfold thickness (V4) has 29.6%, Diastolic blood pressure (V3) has 4.6%, while body mass index (V6), and Plasma glucose concentration (V2) has 1.4% and 0.7% respectively. Number of times pregnant (V1), Diabetes pedigree functions (V7), and Age (V8) have no missing values. Also, the number of valid values are shown along with their means, and standard deviation for all the features with missing values.

Fig. 3 depicts the missing pattern of the features. The features are arranged from left to right with features with non-missing values on the left, to those with least missing values, through the ones with the highest missing values on the right hand. The missing pattern chart displays the value pattern for the analysis function. The pattern represents the group of instances that have the same pattern of incomplete and complete data. In that Fig. 3, Pattern 1 corresponds to instances with no missing value; Pattern 2 shows instances that have missing values on V2; Pattern 3 represents instances with values V6. Pattern 4 represents instances with missing values in V5; Pattern 5 shows instances with missing values on V2 and V5; Pattern 6 represents the missing values on V6 and V5, Pattern 7 is for missing values in V3 and V5; Pattern 8 represents instances with missing values on V4 and V5; Pattern 9 represents instances with missing values on V6, V4, and V5; Pattern 10 is for missing values on V3, V4, and V5; Pattern 11 is for missing values on V6, V3, V4, and V5. All the patterns show no missing values in V1, V7, V8, and V9. Although the dataset has the potential for 2^9 patterns [Lichman, 2013], only 11 feasible patterns are represented in 768 instances.

The features and patterns are arranged in an orderly manner to reveal the existence of monotonicity in the dataset. From the result of the patterns in Fig. 3, it is clear that the missingness in the dataset is non-monotone because all missing cells and non-missing cells are not contiguous; that is, the dataset is MAR as explained in section 3. There are many values to be imputed to achieve monotonicity. Therefore, the use of a monotone (single) method of imputation may not be plausible; the use of multiple imputation techniques (in section 4iii) becomes the needed option.

Multiple imputation techniques were carried out on the dataset features with missing data using (4). The five imputed features were V2, V3, V4, V5, and V6. The order of the imputations of the features was V2, V6, V3, V4, and V5 (in increasing order of the percentage of missingness). The imputation was complete for all missing values in each of the features, and there was no one that was omitted either as a result of 'too' many missing values or no missing value. The descriptive analysis of the imputation on each feature is shown in Tables 2-6.

Table 2 shows multiple imputations for feature V2. Five (5) out of 768 instances in the feature are

missing. The missing values were imputed in 5 iterations, which is a total of 25 runs. The missing values were 100% imputed. The mean, standard deviation, minimum, and maximum values for the original data set and each imputation run is shown in its respective column. The same treatment is done for V3, V4, V5, and V6 in tables 3-6, respectively, in the appendix. The only variation is the number of missing values in each variable which brings about these numbers of imputed values: V3 with 35 missing values has 175 imputed values, V4 with 227 has 1135 imputed values, V5 with 374 missing values has 1870 imputations, and V6 with 11 missing values has 55 imputations.

Observing the characteristics of the datasets before and after the imputations from Tables 2-6, the statistical mean, standard deviation, minimum and maximum are more stable (not at much variance) after the imputation than the reduced datasets during imputation. For example, during imputation in Table 2, the mean imputations for the five iterations are 125.52, 119.22, 145.21, 111.99, 135.09; and the mean imputations for the five iterations of complete datasets are 121.70, 121.67, 121.84, 121.62 and 121.77 which are closed to that of the original data. This shows that the imputed datasets are more reliable than the original form, especially in medical diagnosis, which deals with the issue of saving lives.

The results of the simulated datasets are shown in Fig. 4. The results of the ELM classification of the complete and incomplete datasets show the validation of multiple imputations of Pima India diabetes datasets. P0, the original uncomplete dataset has the least percentage accuracy of 63.0431, while all other five (P1-P5) imputed datasets perform better than the P0. This proof that multiple imputations are a better choice of imputation technique for a non-monotone missing dataset for better classification accuracy.

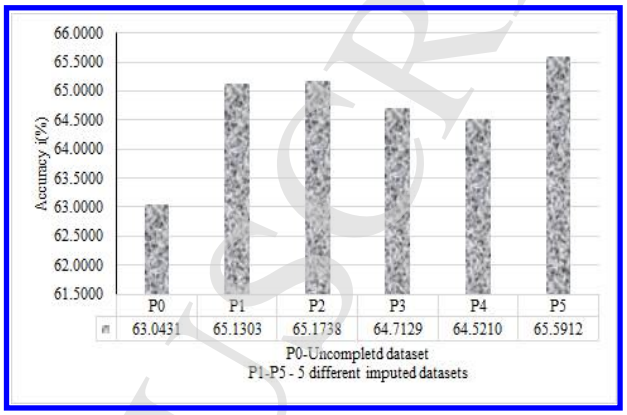

Fig. 2: Extreme learning machine (ELM) validation of multiple imputations on Pima Indian Diabetes showing that the results of classification of imputed datasets (P1-P5) are more accurate than the uncompleted dataset (P0)

8. Conclusion

This paper proposed the examination of the effect of characteristics of missing data on the choice of imputation technique. The mechanism of missingness and the missing pattern were examined as the bases for a choice of imputation technique for filling missing data in the medical dataset with missing values. In this study, we considered the various mechanism for missing values - MAR, MCAR, and MNAR. Different treatments of the missing data based on the mechanisms of missingness were discussed. We backed up our study with investigations into the pattern of missingness in the Pima Indians Diabetes dataset. The observed pattern of missingness on the dataset showed that multiple imputations are more suitable to impute the missing values because it reflects the uncertainty of the undelaying missing data, and the imputations were so treated as shown in Tables 2-6. Our further research work will focus on the performance comparison of different classifiers on the imputed datasets, and suitable optimization technique on a favored classifier in order to improve the accuracy of classification. The work can, however, be extended to compare the accuracy of the imputed datasets with the original dataset with different classifiers like support vector machine (SVM), radial basis function (RBF), and extreme learning machines (ELM).

Acknowledgments

This research has been funded by Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876 and the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under Ministry of Education Malaysia.

References

Lichman, M. [2013] UCI machine learning repository, University of California, School of information & Computer Science.

García-Laencina, P. J., Sancho-Gómez, J. L. and Figueiras-Vidal, A. R. [2010] Pattern classification with missing data: a review. *Neural Comput Appl* **19**, 263–282. doi: 10.1007/s00521-009-0295-6

Tran, C. T., Zhang, M., Andreae, P. *et al*. [2017] An ensemble of rule-based classifiers for incomplete data. In: 2017 21st Asia Pacific Symp. *Intell. Evol. Syst.* IEEE, pp 7–12

Baraldi, A. N. and Enders, C. K. [2010] An introduction to modern missing data analyses. *J Sch Psychol* **48**, 5–37. doi: 10.1016/j.jsp.2009.10.001

Gao, H., Liu, X. W., Peng Y. X. and Jian, S. L. [2015] Sample-Based Extreme Learning Machine with Missing Data. *Math Probl Eng* **2015**,1–11. doi: 10.1155/2015/145156

Gimpy, M. D. R. V. [2014] Missing Value Imputation in Multi-Attribute Data Set. *Int J Comput Sci Inf Technol* **5**, 5315–5321.

Newgard, C. D. and Lewis, R. J. [2015] Missing Data. Howto Best Account for What Is Not Known. *Jama* **314**, 940. doi: 10.1001/jama.2015.10516

Zhu, X. [2014] Comparison of Four Methods for Handling Missing Data in Longitudinal Data Analysis through a Simulation Study. *Open J Stat* **4**, 933–944. doi: 10.4236/ojs.2014.411088

Joseph, J. [2016] How to Treat Missing Values in Your Data. In: *Data Sci. Cent. Online Resour*ces Bid Data Pract. https://www.datasciencecentral.com/profiles/blogs/how-to-treat-missing-values-in-your-data-1. Accessed 14 Feb 2018

Kenward, M. G. [2013] The handling of missing data in clinical trials. **3**, 241–250.

Gautam, C. and Ravi, V. [2015] Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing* **156**, 134–142. doi: 10.1016/j.neucom.2014.12.073

Alade, O. A., Selamat, A. and Sallehuddin, R. [2018] *Recent Trends in Information and Communication Technology*. doi: 10.1007/978-3-319-59427-9

Huang, G. [2015] What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle. *Cognit Comput* **7**,263–278. doi: 10.1007/s12559-015-9333-0

Subbulakshmi, C. V. and Deepa, S. N. [2015] Medical dataset classification: A machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier. Sci World J. doi: 10.1155/2015/418060

Bai, B. M., Mangathayaru, N. and Rani, B. P. [2015] An Approach to Find Missing Values in Medical Datasets. In: *Proc. Int. Conf. Eng.* MIS 2015 - ICEMIS '15. pp 1–7.

Armina, R., Mohd, Z. A., Ali, N. A. and Sallehuddin, R. [2017] A Review on Missing Value Estimation Using Imputation Algorithm. *J Phys Conf Ser.* doi: 10.1088/1742-6596/892/1/012004

Sovilj, D., Eirola, E., Miche, Y., et al. [2015] Extreme learning machine for missing data using multiple imputations. *Neurocomputing* **174**, 220–231. doi: 10.1016/j.neucom.2015.03.108

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

Austin, P. C. and Escobar, M. D. [2005] Bayesian modeling of missing data in clinical research. *Comput Stat Data Anal* **49**, 821–836. doi: 10.1016/j.csda.2004.06.006

Falcaro, M. and Carpenter, J. R. [2017] Correcting bias due to missing stage data in the non-parametric estimation of stage-specific net survival for colorectal cancer using multiple imputations. *Int J Cancer Epidemiol Detect Prev* **48**, 16–21. doi: 10.1016/j.canep.2017.02.005

Tang, J., Zhang, G., Wang, Y., *et al* [2015] A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transp Res Part C Emerg Technol* **51**, 29–40. doi: 10.1016/j.trc.2014.11.003

Nguyen, C. D. and Carlin, J. B. and Lee, K. J. [2017] Model checking in multiple imputation: An overview and

case study. *Emerg Themes Epidemiol* **14**, 1–12. doi: 10.1186/s12982-017-0062-6

Liu, Y., Gopalakrishnan, V. [2017] An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. *Data* **2**, 8. doi: 10.3390/data2010008

Diaconis, P. and Efron, B. [1983] Computer intensive methods in statistics. Sci Am 248:115–130.

Huang, G. B. and Chen, L. [2007] Convex incremental extreme learning machine. *Neurocomputing* **70**, 3056–3062. doi: 10.1016/j.neucom.2007.02.009

Shang, Z. and He, J. (2015) Confidence-weighted extreme learning machine for regression problems. Neurocomputing 148:544–550. doi: 10.1016/j.neucom.2014.07.009

Wasito, I. and Mirkin B (2006) Nearest neighbors in least-squares data imputation algorithms with different missing patterns. *Comput Stat Data Anal* **50**, 926–949. doi: 10.1016/j.csda.2004.11.009

Mukaka, M., White, S. A., Terlouw, D. J., *et al* [2016] Is using multiple imputations better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials* **17**, 1–12. doi: 10.1186/s13063-016-1473-3

Zhang, Z. [2016] Missing data imputation: focusing on single imputation. *Ann Transl Med* **4**, 9. doi: 10.3978/j.issn.2305-5839.2015.12.38

Gelman, A. and Hill, J. [2007] *Data Analysis Using Regression and Multilevel/Hierarchical Models*, **1** ed. Cambridge University Press, New York

Rodwell, L., Lee, K. J., Romaniuk, H and Carlin, J. B. (2014) Comparison of methods for imputing limited-range variables: A simulation study. *BMC Med Res Methodol* 14:1–11. doi: 10.1186/1471-2288-14-57

van Kuijk, S. M. J., Viechtbauer, W., Peeters, L. L. and Smits L. [2016] Bias in regression coefficient estimates when assumptions for handling missing data are violated: A simulation study. Epidemiol Biostat Public Heal 13:1–8. doi: 10.2427/11598

Smith, J. W., Everhart, J. E., Dickson, W. C., *et al* [1988] Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proc Annu Symp Comput Appl Med Care* 261–265.

Strack, B., Deshazo, J. P., Gennings, C., *et al* [2014] Impact of HbA1c Measurement on Hospital Readmission Rates : Analysis of 70 , 000 Clinical Database Patient Records. *Biomed Res. Int.* 2014