A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms

T. Makaba[1], E. M. Dogo[2]
[1] Department of Applied Information Systems,
[2]Department of Electrical and Electronic Engineering Science, Institute for Intelligent Systems
University of Johannesburg
Johannesburg, South Africa
[1]tmakaba@uj.ac.za, [2]eustaced@uj.ac.za

***Abstract*— Dealing with missing values in data is an important feature engineering task in data science to prevent negative impacts on machine learning classification models in terms of accurate prediction. However, it is often unclear what the underlying cause of the missing values in real-life data is or rather the missing data mechanism that is causing the missingness. Thus, it becomes necessary to evaluate several missing data approaches for a given dataset. In this paper, we perform a comparative study of several approaches for handling missing values in data, namely listwise deletion, mean, mode, k–nearest neighbors, expectation-maximization, and multiple imputations by chained equations. The comparison is performed on two real-world datasets, using the following evaluation metrics: Accuracy, root mean squared error, receiver operating characteristics, and the F1 score. Most classifiers performed well across the missing data strategies. However, based on the result obtained, the support vector classifier method overall performed marginally better for the numerical data and naïve Bayes classifier for the categorical data when compared to the other evaluated missing value methods.**

Keywords - missing data; imputation methods; performance metrics; machine learning, classification

I. INTRODUCTION

Approaches to dealing with missing data have been well researched in literature, using either statistical [1], [2] or computational intelligence (such as machine learning (ML)) [3], [4] approaches. Missing values in data are broadly categorized into three missingness mechanisms [1], [2]: data missing completely at random (MCAR) when the probability of an instance case or variable having a missing value is not dependent on either the known value itself or any other value or variable in the given dataset; data missing at random (MAR) when the probability of an instance or variable having a missing value is dependent on other known variables but not on the value of the missing data itself; data missing not at random (MNAR) when the probability of an instance or variable having a missing value is dependent on the value of that variable itself. Missing data are now a common problem in many real-world datasets in numerous domains such as fraud detection, sensor readings, anomaly detection etc. The missingness can be attributed to numerous sources and reasons such as measurement error, mechanical faults, non-response or deleting of values [5]. Missing data, if not addressed during the data preprocessing stage prior to feeding these into an ML model, could induce complexity into the data analysis and affect the performance of ML algorithms in terms of conclusions that can be inferred from the data, because of reduced data samples and bias in estimation of the algorithms' parameters. Numerous missing data imputation handling techniques have been developed [6], which could be broadly categorized as listwise or case deletion, single and multiple imputations. Researchers continue to develop enhanced variants. On the other hand, some researchers have carried out a comparative evaluation of the current missing data techniques to provide more insight and guidance on the choice of techniques, depending on the percentage, pattern and mechanism underlining the missingness in a dataset [3], [5], [7]-[10].

This study compares six missing data-handling methods, namely, listwise deletion (LD), mean, mode, k-nearest neighbor (k-NN), expectation-maximization single imputation (EMSI) and multiple imputations by chained equation (MICE), on six ML algorithms: logistic regression (LR), k-NN, support vector machine (SVM), random forest (RF), naïve Bayes (NB) and artificial neural network (ANN). Two real-life datasets are used and evaluated based on the following performance metrics: accuracy, root mean squared error (RMSE), receiver operator characteristics (ROC) and the F1-score.

The rest of the paper is organized in the following way: Section II reviews the literature with regard to the missing values and imputation strategies and the classifiers employed in this study. Section III outlines the study methodology, which comprises the experimental set-up, data set used, and the performance metrics for evaluation. Section IV provides the results achieved and a discussion on these. Finally, section V concludes the paper.

II. MISSING DATA METHODS

The term missing data refers to the absence of records or values or observations usually expected to be present in a dataset. Missing data strategies are broadly categorized into three: (1) filling with zero, or ignoring data with missing values, or deleting or dropping missing values, (2) single imputation strategies and (3) multiple imputation strategies. Four of the methods used in this study are based on single imputation, while one is based on multiple imputation methods (IM). The methods that are considered in this study are briefly described as follows:

A. Listwise Deletion

LD is a statistical method that handles missing data by deleting or ignoring the entire record of missing values in a dataset, and thus excluding these from the analysis. Only the complete data are retained, which can result in biased estimations. This method is also referred to as complete-case analysis and assumes that data are MCAR [8].

B. Imputation Methods

Imputation is an approach to handling missing data by estimating the missing values in a dataset. The IM could be subdivided into single and multiple IM. The methods considered in this paper are briefly described as follows:

1) Mean/Mode: Mean consists of replacing the missing data for a given variable by the mean or mode of all known values of that variable. Generally, the mean method is suitable for numerical variables and the mode for categorical variables. Mean or mode usually assumes MCAR [1].

2) k-Nearest Neighbors: k-NN defines a set of k-NNs for each sample or individual and then replaces the missing data for a given variable by averaging through estimating (non-missing) values of its neighbors. The size of the dataset to be analyzed and the optimal k value are crucial for this method. k-NN usually assumes data are MCAR [8].

3) Expectation maximization (EM): EM is an iterative means of imputing one or more plausible missing data (EM single or multiple imputations) values, resulting a complete new dataset, through a repeated procedure [2], [11]. EM usually assumes that data are MAR.

4) Multiple imputations by chained equations: The MICE method is an iterative algorithm based on chained equations that use an imputation model specified separately for each variable and involving the other variables as estimators. MICE is a multiple imputation method that involves imputing missing values in a dataset not once, but many times [1]. MICE usually assume that data are MAR.

The criteria and justification for choosing of missing data methods are based on their popularity and how often they have been cited and used in literature, as suggested in Table 1.

III. MACHINE LEARNING MODELS

The six classifiers are selected based on their different forms of learning methods. This ensures a broader consideration of families of algorithms depending on their learning philosophies: linear, density-based models, instance-based, tree and neural network-based models [12]. These allow researchers a robust assessment of the missing data methods.

1) Logistic Regression (LR): LR is a linear-based classifier that calculates the linear output, followed by a stashing function over the regression output. LR is an easy, fast and simple ML method.

2) k-Nearest Neighbors: The k-NN classifier is an instance-based method where new instance query results are classified according to the majority k-NN of the category using the Euclidean distance. The basic logic of the k-NN is to explore the nearest neighbor by assigning an initial size of k neighborhood [13]. One of the main advantages of k-NN is that it is an easy and simple ML algorithm.

3) Support Vector Machine: SVM is a supervised ML algorithm that uses a technique called the kernel trick to transform the dataset and from the transformation it finds the best boundary between the possible results.

4) Random Forest (RF): The RF model is an ensemble and tree-based learning method that can be used to build predictive models. It combines a number of decision tree classifiers and averages their predictive accuracy, in the process improving on the overall model performance. Ensemble learning uses multiple learning models to gain better predictive results [12].

5) Naïve Bayes: The NB classifier is a probabilistic learning technique that is based on the Bayes theorem, which assumes features are statistically independent. NBC uses prior knowledge to calculate the probability of a sample for a certain category [12].

6) Artificial Neural Networks: An ANN examines the relationship between inputs and outputs by using the training dataset without much detail about the system; it mimics the workings of the human brain [12].

IV. RELATED WORK

A considerable number of research articles are available to deal with missing values across several domains. Some of the earlier research works focused on developing enhanced missing data IM, such as in [4], while others focused on performing a comparative analysis of existing missing data methods on different ML algorithms, such as in [3], [7], [14]. Most of the articles apply single imputation strategies in dealing with missing values in the dataset, since, it is very often unclear what the underlying causes of missing values in any given data are and hard to know in advance which missing value method is ideal for a given dataset or problem [10]. In addition, applying missing data imputation have is likely to distort variable distribution and associated interactions, and in a way also affects the ML model. It is for this reason that we embark on conducting an experimental comparison of several missing data approaches for our real-world dataset against different ML classification algorithms. In this way we could gain valuable insights into the biases shown by these missing values strategies and how they affect different learning classification algorithms for our given datasets. From the summary of some related works outlined in Table 1, it appears that the following missing data methods are the most popularly used: mean/mode, k-NN, EM and multiple imputations such as MICE.

V. STUDY METHODOLOGY

A. Experimental Set-up

The aim of this experiment was to carry out a comparative analysis and evaluate the impact of five missing data-handling methods against six classifier ML algorithms with four performance metrics using two real-world datasets.

TABLE 1. SUMMARY OF RELATED WORKS

#	Study	Methods/Algorithms	Dataset	Metrics	Findings
1	[3]	Comparison of IM based on ML: MLP, SOM and k-NN with statistical imputation-based methods: mean, hot-deck and multiple imputations (MI) and EM	Breast cancer from *El A´ lamo-I* project in Spain.	ROC curve Friedman's test, Pairwise test	The results of the study showed that ML IMs outperformed statistical IMs when predicting a patient's outcome.
2	[15]	IM: Comparison of six MICE methods.	Iris	Mean confidence interval length and mean standard error	The results of the study revealed that MICE in combination with Bayesian regression produced the least standard error and mean confidence interval length.
3	[4]	IM: Comparison of mean, k-NN and evolutionary k-NN	Gene expression	Mean error	Evolutionary k-NN outperformed the normal k-NN and mean methods
4	[7]	ML: Decision tree (DT) Missing data methods: LD, EMSI, EMMI, Surrogate variable splitting, DT single imputation, mean or more single imputation and fractional cases	Twenty-one UCI ML repository	Excess error	Multiple imputations using EM algorithm represented a superior approach to handle incomplete data.
5	[14]	ML: Bayesian Networks (BN) IM: k-NN	Medical obstructive sleep apnea	ROC, AUC, Sensitivity analysis and specificity	k-NN imputation approach proved a far better solution than LD.
6	[12]	ML: CART, k-NN, LDA, NBC, repeated incremental pruning to produce error reduction (RIPPER), SVM and C4.5.	Gauteng road traffic accident	Error rate and Excess error rate	The proposed tree-based classifier imputation method was evaluated against seven classifiers: C4.5, CART, KNN, LDA, NB, RIPPER and SVM across three missing data mechanisms: MCAR, MAR and IM. The proposed method proved robust and efficient in comparison to existing methods
7	[9]	ML: regression model Six IM: EMMI with bootstrapping, MI using multiple correspondence analysis, MI using latent class analysis, multiple hot-deck, MICE based LR and MICE based RF	Questionnaire-based study in the Norwegian opioid maintenance treatment program	Standard error	MI using multiple correspondence analysis had the best overall performance.
8	[16]	ML: BN IM: EM AND MI	Alarm network	Cross-entropy and log-likelihood	Evaluated Bayesian network on incomplete dataset based on MCAR and MAR; the proposed algorithm performed better compared to commonly used adhoc methods.
9	[5]	Comparison of six IM: mean, k-NN, fuzzy k-means, singular value computation, Bayesian principal component analysis and MICE	Iris, *E. coli* and breast cancer	RMSE, unsupervised classification error, supervised classification error and execution time	bPCA and fKM showed better performance based on the MCAR assumption.
10	[8]	ML: RF, k-NN, ANN and SVM Missing methods: LD, mean-mode, k-NN and regression imputation	Two UCI remote sensing	Accuracy, mean absolute error, RMSE, precision, ROC	k-NN was a better performer with regression imputation, while RF was the worst performer

Our experiments were conducted on 'SPyDER' (*Scientific Python Development EnviRonment*) on Anaconda Python distribution, each time using one missing data method to test the chosen ML algorithms. The experimental simulation is a three-way repeated-measures strategy, which allows the main effect factors (6 classifiers, 6 missing data methods and 4 performance metrics) to be evaluated against interaction with the random effect factor (numerical and categorical) datasets. Throughout the experimentation, we kept the default settings of the presented classifiers. However, for the categorical data, we only considered LD and the most frequent (mode) missing strategies because of the size of the dataset, the number of missing values and our observation with regard to k-NN, EMSI and MICE strategies, which did not show much difference with the numeric dataset, as shown in Table 3.

B. Dataset

The experiments were carried out using two real-life data sets, namely Gauteng road traffic and water quality datasets. The characteristics of the dataset are summarized in Table 2.

TABLE 2. SUMMARY OF DATA SET CHARACTERISTICS USED IN THE EXPERIMENTS

Dataset	Data Type	Instances	Attributes	Class	Missing values	Missing values %
Gauteng road traffic	Nominal categorical	672	4	3	21	3.12
Water quality data	Continuous numerical	1000	9	2	200	20

C. Performance Metrics

The following performance metrics were used to evaluate the performance of the models after implementing the missing data methods: Accuracy, RMSE, ROC and F1-score. The four chosen metrics are the most popular methods used for evaluating classification ML algorithms [17].

VI. RESULTS AND DISCUSSION

Table 3 and Figure 1 show the results for numerical water quality data, while Table 4 and Figure 2 show the result for categorical Gauteng road traffic data. The results report the performance of the examined classifiers based on different missing data methods with a constant percentage of missing values. The following is observed:

With regard to the numerical data, generally all classifiers performed well across the different missing data strategies used in this study. However, overall SVC performed with consistency and slightly better in terms of all the performance metrics evaluated, with the NB classifier showing the marginally lowest performance except when using the LD and mode methods. In addition, LD, mean and mode performed well across all the classifiers compared to the more advanced k-NN, EMSI. The reasons for their performance, apart from ease of implementation, are the low occurrence of missing values in the numerical dataset and variance reduction. Moreover, we observed that the MICE method performed well for all the classifiers. One possible reason is that it takes into account the uncertainties resulting from guesses created by other IM, by taking into cognizance all the available information from other variables in the data and averaging their results for better estimates of the unknown true missing value. It could thus provide more valid standard errors, p-values and final inferences. However, computational cost is one of MICE's drawbacks.

With regard to the categorical data, overall NBC seems to perform slightly better on both LD and mode strategies used in comparison to the other classifiers. One reason for this is that generally, NBC performs well with a smaller dataset with a low missing rate. On the other hand, ANN had the lowest RMSE for the LD and mode methods in comparison to all the other classifiers, indicating better fit of ANN model and classification accuracy. However, all the classifiers examined performed slightly better against the mode strategy in comparison to the LD method. Because data are lost when using the LD method, complexity could be added in term of variance and bias. In general, we observed that the results obtained varied depending on the classifier, type of data (numerical or categorical), and percentage of missing value. This means that no single missing data methods is superior or fits all dataset type problems. We have seen in our case that results varied with both numerical and categorical datasets, reasons such as how correlated the attributes are, the data distribution pattern, data size, missing value rate and data type. Different missing value methods induce biases, particularly if the methods are based on certain assumptions pointed out earlier in section II.

TABLE 3. RESULTS OF DATASET 1 (Numerical)

Models/Metrics	**Missing Data Methods**					
	LD	***Mean***	***Mode***	***k-NN***	***EMSI***	***MICE***
Accuracy						
LR	1.00	1.00	1.00	1.00	0.975	1.00
k-NN	1.00	1.00	1.00	0.995	0.980	1.00
SVC	1.00	1.00	1.00	0.995	1.00	1.00
NB	1.00	0.97	1.00	0.980	0.91	0.97
RF	1.00	1.00	1.00	1.00	0.995	1.00
ANN	1.00	1.000	1.00	1.00	0.985	1.00
RMSE						
LR	0.0010	0.007	0.00033	0.00192	0.0233	0.0062
k-NN	0.00	0.000	0.000	0.005	0.0148	0.00
SVC	0.000059	0.000075	0.000117	0.00257	0.000654	0.000075
NB	0.00	0.029	0.000	0.0142	0.0766	0.0286
RF	0.000313	0.00030	0.0004	0.00045	0.0041	0.0003
ANN	0.000467	0.000342	0.000045	0.000598	0.0556	0.000648
ROC						
LR	1.00	1.00	1.00	1.00	0.995	1.00
k-NN	1.00	1.00	1.00	0.9957	0.991	1.00
SVC	1.00	1.00	1.00	1.00	1.00	1.00
NB	1.00	1.00	1.00	0.998	0.973	1.00
RF	1.00	1.00	1.00	1.00	1.00	1.00
ANN	1.00	1.00	1.00	1.00	0.9957	1.00
F1-score						
LR	1.00	1.00	1.00	1.00	0.971	1.00
k-NN	1.00	1.00	1.00	0.994	0.977	1.00
SVC	1.00	1.00	1.00	0.994	1.00	1.00
NB	1.00	0.96	1.00	0.976	0.854	0.963
RF	1.00	1.00	1.00	1.00	0.994	1.00
ANN	1.00	1.00	1.00	1.00	0.9825	1.00

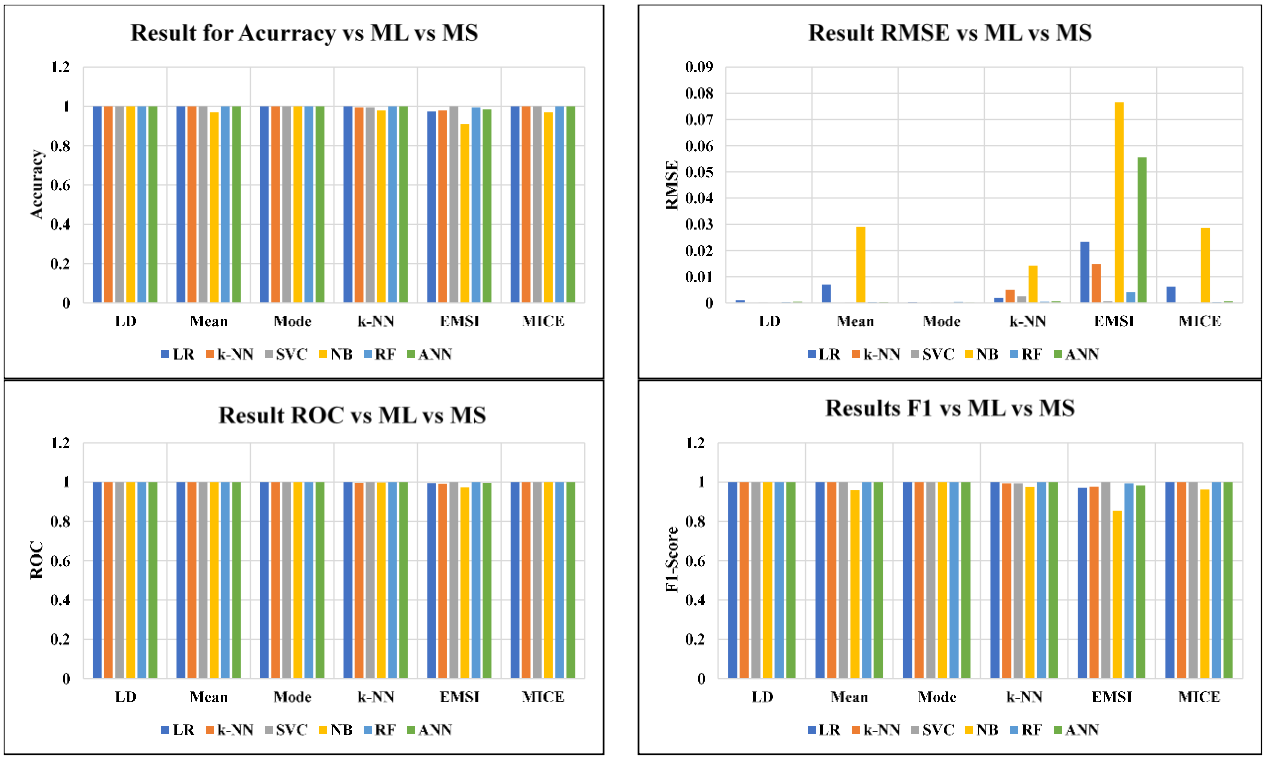

Fig. 1.Performance result ML vs MS on Dataset 1 *(Numerical)*

TABLE 4. RESULTS DATASET 2 (Categorical)

Models/Metrics	**Missing Data Methods**	
	LD	***Mode***
Accuracy		
LR	0.885	0.911
k-NN	0.863	0.911
SVC	0.878	0.896
NB	0.90	0.911
RF	0.879	0.896
ANN	0.86	0.90
RMSE		
LR	2.81	2.68
k-NN	2.84	2.66
SVC	2.72	2.56
NB	2.70	2.58
RF	2.81	2.68
ANN	0.14	0.10
ROC		
LR	0.92	0.96
k-NN	0.83	0.92
SVC	0.96	0.96
NB	0.96	0.96
RF	0.87	0.94
ANN	0.88	0.92
F1-score		
LR	0.89	0.92
k-NN	0.86	0.91
SVC	0.88	0.90
NB	0.91	0.92
RF	0.88	0.90
ANN	0.86	0.90

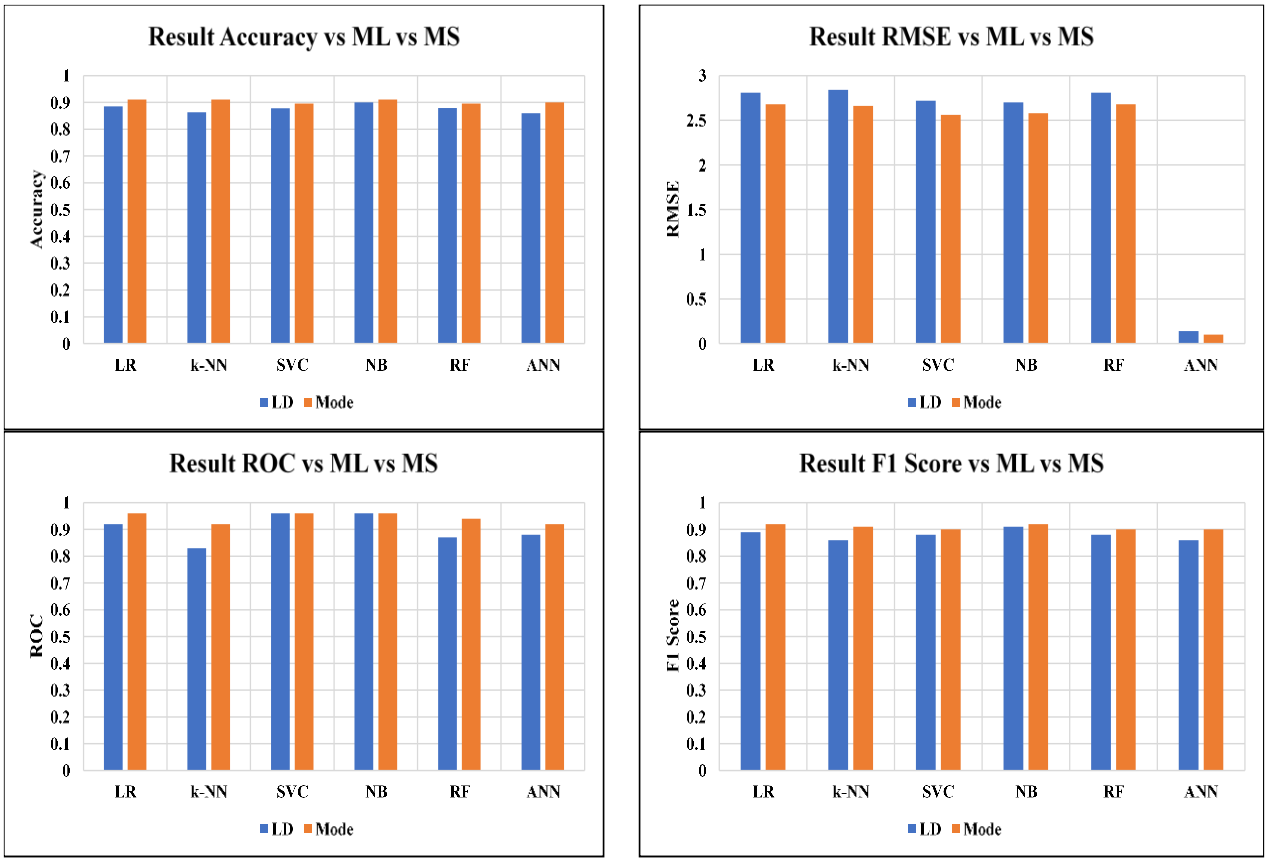

Fig. 2. Performance result ML vs MS on Dataset 2 *(Categorical)*

VII. CONCLUSION

The aim of this work was to evaluate the performance of six ML classifiers on different missing data strategies using numerical and categorical datasets. We observed a very marginal difference in terms of overall performance across all the classifiers. However, SVC performed marginally better for the numerical dataset, while NB classifier did the same for the categorical dataset across the missing data methods examined. However, ANN had the lowest RMSE when compared to all the other classifiers for the categorical dataset, indicating better fit of ANN model. Nonetheless, for the categorical dataset, we noticed slightly improved performance by the classifiers against mode method in comparison to the LD method. We intend to test other missing value strategies, including ML and missing data methods in the future, using larger datasets and different missing values rates. The authors would like to pay detailed attention to employing ML approaches to handling missing data, statistical quantification of biases and sensitivity analysis for the missing data strategies as areas of interest in future work. Finally, our preliminary submission is that knowing the cause of missing values in a dataset is key to tackling the missingness problem, since the missing value methods are based on certain assumptions.

ACKNOWLEDGEMENT

We would like to thank the University of Johannesburg for funding and making the resources available to complete this work. The authors are also thankful to Mikros Traffic Monitoring (Pty) Ltd and Prof. T. Bartz-Beielstein for making the datasets available.

REFERENCES

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Third Edition (eds R. Little and D. Rubin), 2019. DOI: 10.1002/9781119482260

[2] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, *(3)*, pp. 581-592, 1976.

[3] J. M. Jerez, L. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, *(2)*, pp. 105-115, 2010. Available: https://www.clinicalkey.es/playcontent/1-s2.0-S0933365710000679. DOI: 10.1016/j.artmed.2010.05.002.

[4] H. de Silva and A. S. Perera, "Missing data imputation using evolutionary k- nearest neighbour algorithm for gene expression data." Sep 2016. Available: https://ieeexplore.ieee.org/document/7829911. DOI: 10.1109/ICTER.2016.7829911.

[5] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 6, *(1)*, 2015. DOI: 10.4172/2155-6180.1000224.

[6] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, *(5)*, pp. 402-406, 2013. Available: http://synapse.koreamed.org/search.php?where=aview&id=10.4097/kjae.

2013.64.5.402&code=0011KJAE&vmode=FULL. DOI: 10.4097/kjae.2013.64.5.402.

[7] B. Twala, "An empirical comparison of techniques for handling incomplete data using decision trees," *Applied Artificial Intelligence,* vol. 23, *(5),* pp. 373-405, 2009. Available: http://www.tandfonline.com/doi/abs/10.1080/08839510902872223. DOI: 10.1080/08839510902872223.

[8] T. Nkonyana and B. Twala, Eds., *Impact of Poor Data Quality in Remotely Sensed Data.* (Artificial Intelligence and Evolutionary Computations in Engineering Systems ed.) Singapore: Springer Nature.

[9] M. R. Stavseth, T. Clausen and J. Røislien, "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data," *SAGE Open Medicine,* vol. 7, pp. 2050312118822912, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/30671242.

[10] T. Marwala, *Computational Intelligence for Missing Data Imputation, Estimation, and Management.* 2009Available: https://ebookcentral.proquest.com/lib/[SITE_ID]/detail.action?docID=3309570.

[11] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the Em algorithm," *Journal of the Royal Statistical Society,* vol. 39, *(1),* pp. 1-38, 1977. Available: http://www.econis.eu/PPNSET?PPN=388257237.

[12] B. Twala and F. Mekuria, "Ensemble multisensor data using state-of-the-art classification methods." Sep 2013. Available: https://ieeexplore.ieee.org/document/6757711. DOI: 10.1109/AFRCON.2013.6757711.

[13] B. Twala, "Dancing with dirty Road traffic accidents data: The case of Gauteng Province in South Africa," *Journal of Transportation Safety & Security,* vol. 4, *(4),* pp. 323-335, 2012. Available: http://www.tandfonline.com/doi/abs/10.1080/19439962.2012.702711. DOI: 10.1080/19439962.2012.702711.

[14] D. Ferreira-Santos, M. Monteiro-Soares and P. P. Rodrigues, "Impact of imputing missing data in Bayesian network structure learning for obstructive sleep apnea diagnosis," *Studies in Health Technology and Informatics,* vol. 247, pp. 126-130, 2018. Available: https://www.ncbi.nlm.nih.gov/pubmed/29677936.

[15] G. Chhabra, V. Vashisht and J. Ranjan, "A comparison of multiple imputation methods for data with missing values," *Indian Journal of Science and Technology,* vol. 10, *(19),* pp. 1-7, 2017. DOI: 10.17485/ijst/2017/v10i19/110646.

[16] M. Singh, "Learning Bayesian Networks from Incomplete Data," *In AAAI/IAAI,* pp. 539., 1997.

[17] C. Ferri, J. Hernndez-Orallo and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recog. Lett.,* vol. 30, *(1),* pp. 27-38, 2009.