2019 IEEE 4th International Conference on Computer and Communication Systems

Application of the Modified Imputation Method to Missing Data to Increase Classification Performance

Elenita T. Capariño
Graduate Programs
Technological Institute of the Philippines
Quezon City, Philippines
e-mail: elenita.caparino@bulsu.edu.ph

Ariel M. Sison
School of Engineering and Technology
Emilio Aguinaldo College
Manila, Philippines
e-mail: ariel.sison@eac.edu.ph

Ruji P. Medina
Graduate Programs
Technological Institute of the Philippines
Quezon City, Philippines
e-mail: ruji.medina@tip.edu.ph

***Abstract*—Incomplete data or missing data diminishes the effectivity of statistical results, and may cause bias estimates, which in turn leads to unsound judgment. Inefficiency and impediments in data treatment analysis, which are among the predicaments linked with missing values, may affect the supervised learning process and reduce the classification accuracy and performance of the prediction model in a data mining task. This study applied the modified imputation method—which was previously tested with well-known imputation algorithms—to renowned classification techniques namely Naïve Bayes, One-R, k-Nearest Neighbor (kNN), C4.5, and Support Vector Machine (SVM) using open data sets from the UCI Repository. The level of performance in terms of precision, accuracy, and Receiver Operating Characteristics (ROC) using Weka tool, before and after imputation was examined. This study manifests that there was an improvement in the classification performance upon the application of the modified imputation method on datasets during preprocessing, compared to that of datasets with missing values.**

Keywords-classification; data mining; imputation; missing values; machine learning

I. INTRODUCTION

Missing values may conceal essential knowledge about a dataset. Wresting knowledge from a pool of data with missing values may create serious problems, especially when data mining algorithms have been applied. Most methods used in forecasting are nearest neighbor and Naïve Bayesian classifiers, which cannot handle data that contain missing values [1]. Data sets in the real world are vulnerable to incomplete data or missing values and may result in bias; affect supervised learning process, weaken the accuracy of data analysis and classification algorithms, and reduce data value [2] [3].

Incomplete or missing data diminishes effectivity of statistical results and may cause biased estimates, which, in turn leads to unsound judgment [4]. Predicaments linked to missing values in data mining tasks are: inefficiency; impediments in data treatment and analysis and bias. [5]

Missing values exist in almost all researches, and this may present a degree of vagueness in any data analysis. Therefore, researchers have to take them into account and treat them appropriately in order to arrive at an effective and acceptable data analysis. [6]

Accuracy of prediction models may be degraded in the presence of missing data. Several studies affirmed that performance accuracy of classification algorithms in a data mining process tends to decline as missing data percentage increases. [7] [8]

The goal of any classification task is to predict a target class in each data instance accurately. In the same way, the main objective of a classification task is to yield results that are more precise, definite, and accurate. [9] [10]

Hence, this study aims to increase performance of classification algorithms by handling missing values and applying the modified imputation method to data sets before and after imputation. In this study, C4.5, One-R, Naïve Bayes, kNN and Support Vector Machine were the classification methods used after applying the modified imputation method.

II. RELATED WORKS

Missing data affects the accuracy of prediction models in a classification task of a data mining process. A study was conducted which reveals the effect of missing data to well-known classifiers as the sensitivity of six classification algorithms to missing data were analyzed. Among these classifiers were Naïve Bayes, Logistic Regression, Neural Network using backpropagation technique, kNN, C4.5, and Logistic Model Trees. They ascertained that in a data set with an increasing rate of missing values, accuracy of the classifier tends to decrease [7]. Similarly, in another study, the impact of missing values was examined in several well-known data mining classifiers; and authors determined that accuracy performance of the prediction model was loss as the missing data percentage increases. [8]

978-1-7281-1322-7/19/$31.00 ©2019 IEEE

```
Input:    Dataset with missing values

Step 1.  Determine SpearmanCorrelation on all attributes
Step 2.  Pair highest correlation attribute with missing values' attribute
Step 3.  Perform the following:
    3.1  X = attribute with missing values
    3.2  Y = highest correlation attribute
    3.3  cnt = count of missing values
    3.4  k=1
    3.5  While k!=cnt:
            3.5.1  Perform Impute_Routine
            3.5.2  Increment k
    3.6 Output Ximputed Values
    3.7 If there exists other attributes with missing values then
            3.7.1  Reset X, Y,cnt, k
            3.7.2  Process the next attribute with missing values
            3.7.3  Repeat Step 3.

Output:  Complete data
```

Figure 1. The main routine of the modified imputation method.

There were also studies on creating imputation schemes to address specific applications. In one study, an imputation method for pattern classification was proposed, where the major goal is to reduce bias caused by the absence of information. The authors examined the performance of their method by using three algorithms which represents three groups of classification technique: 1) rule induction learning, 2) approximate models and 3) lazy learning. The classifiers used were C4.5, Naïve Bayes, and kNN respectively and adapted the Wilcoxon signed-rank test in evaluating the imputation scheme [11]. Similarly, another imputation method was proposed for text mining applications which was applied and investigated using only C4.5 classifier; however, the comparison with other imputation methods was not performed. [12]

In another study, classifier and imputation scheme were paired based on the properties of missing data in datasets [13]. Correspondingly, an algorithm for missing data imputation was also proposed, which is appropriate for "mixed-attributes" datasets. Three groups of classifiers were used: C4.5, One-R, Naïve Bayes, and KNN. [14]

There are a number of performance measures for classification tasks. Confusion matrix is one, which is a widely-used matrix for its function of evaluating classification model performances. Accuracy, precision, recall, and F-measure were the standard metrics which can be derived from the confusion matrix [15]. It is a popular means to demonstrate the accuracy of the solution to a classification task. Data in the matrix illustrates the actual and predicted values performed by a classification model, and it indicates the correct and incorrect predictions contrasted to the true target value. [16]

Performance which is correspondent estimation-oriented measure focuses on the rows which contains estimated classes on the confusion matrix. One of which is the Positive Predictive Value (PPV) which is otherwise known as Dice's association index, user's accuracy, or Precision. Precision is observed as a metric to measure relevance, quality and exactness. Precision is based on a fraction of retrieved instances which are relevant and is one of the effective performance metrics used in most classification tasks. Several researches have applied precision as performance measure in classification process. [17] [18] [19]

```
Input: attribute with missing values as X

1. To fill up the missing values of X, use the values of Y to match it to other records.
2. For each missing Xi, tag its variable Y value as Yi.
   (i=1 to k=count of missing X's)
3. If Yi is also missing then the imputed value of Xi will be the MEAN of ALL X's.
4. If Yi is not missing, then proceed to next step.
5. Get the minimum absolute difference of Yi to other Y's.
   CASE 1
      If minimum absolute difference of Yi from Yj is ZERO

        If UNIQUE
          If Xj value is not missing then imputed value is Xj.
          If Xj is missing, proceed to Case 2 using the next min abs difference.
        If NOT UNIQUE
          If at most one is not missing then imputed value is Xj
          If at least 1 is not missing then imputed value is MEAN of ALL Xj's
          If all are missing, proceed to Case 2 using the next min abs difference.
   CASE 2
      If minimum absolute difference of Yi from Yj is NOT ZERO
        If UNIQUE
          If Xj value is missing then imputed value is MEAN of ALL X's
          If Xj value is not missing then imputed value is Xj
        If NOT UNIQUE
          If at most one is not missing then imputed value is Xj
          If at least 1 is not missing then imputed value MEAN of ALL Xj's
          If all are missing then imputed value is MEAN of ALL X's
6. If X is integer and the imputed Xj is not, round it using standard rules of
   rounding.
7. Repeat until Xk has been imputed.

Output:    Imputed values of attribute X
```

Figure 2. Pseudocode of the modified imputation routine.

Another alternative for evaluation is being used—the AUC or "the area under the ROC (Receiver Operating Characteristics) curve". ROC analysis, otherwise known as Receiver operating characteristics (ROC) analysis graphically depicts a prediction performance, and it also demonstrates the tradeoffs between specificity and sensitivity. [18]

This study compares the classification performance using accuracy, precision and ROC values before and after the modified imputation method was applied to three given datasets. Findings and results are discussed in the later sections of this paper.

III. THE MODIFIED IMPUTATION METHOD

Fig. 1 is the pseudocode of the main routine of the modified imputation method. The input required for this is the dataset containing missing values.

At first, Spearman correlation test was determined on all attributes. After which, obtain and pair the highest correlated variables; that is, variable X is the attribute with missing values, and variable Y is the attribute which will be used in imputing the missing values.

The Spearman correlation formula:

$$(R) = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (1)$$

where d is the difference between ranks, and n is the instance.

Next is the Pearson correlation, which is shown below, where X and Y are the variables, and n is the instance.

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n}(\sum Y^2 - \frac{(\sum Y)^2}{n})}} \quad (2)$$

From here, imputation routine will be performed for every attribute with missing values. Fig. 2 shows the imputation routine which will be executed for each attribute with missing values; and this will be the source where the new values will be derived, which will replace the missing ones on each iteration.

Coming from the main routine, K is the count of missing values for a certain attribute. The absolute value difference was then determined for attribute Y. Assume j as the number of instance, X_1 is missing, Y_1 has a value, the absolute value difference is determined using:

$$a_1 = |Y_1 - Y_1|, a_2 = |Y_1 - Y_2|, \ldots a_j = |Y_1 - Y_j| \quad (3)$$

Unique minimum difference absolute value shall be observed. If the minimum difference is not unique, then the mean of the X attributes with the minimum difference shall be computed. And this will be the imputed value. The formula for mean is shown below:

$$Mean = \frac{\sum X}{n} \quad (4)$$

where X are the values of the attributes, and n are the number of instances.

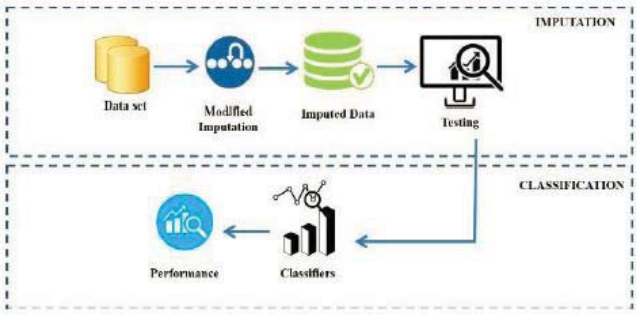

Figure 3. Framework of the study.

If both X and Y values are missing, then the mean of all the X attributes will be the imputed value. Round off X imputed if variable X is integer.

IV. METHODOLOGY

Fig. 3 depicts the framework of the study. The study was subdivided into two major phases namely the Imputation Phase, and the Classification Phase. On the first phase, an imputation method was devised. Prior to this paper, the modified imputation method (MODIMP) was tested with well-known imputation schemes namely mean imputation, regression, expectation maximization, and Markov Chain Monte Carlo; using open data sets with stricken missing values [20]. The performance measures were root mean square error, bias, mean square error and execution time. Results showed that MODIMP performs better among these imputation techniques, and ranked second after expectation maximization method. In this paper, the modified imputation method was applied to well-known classification algorithms which were classified into three, namely rule induction, lazy learning and approximate models, and these were Naïve Bayes (NB), One-R, k-Nearest Neighbor (kNN), C4.5 and Support Vector Machine (SVM). Classification performances were contrasted before and after imputation on two (2) data sets with applied stricken missing values (5%, 10%, 15%, 20%, 25% and 30%), and one (1) data set which originally has missing observations. Table I shows the data sets used in this study.

TABLE I. DATA SETS

Dataset	Instances	Attributes	Missing Values
Wine	178	14	Applied 5%,10%, 15%, 20%, 25%, 30%
Iris	150	5	Applied 5%,10%, 15%, 20%, 25%, 30%
Breast cancer	699	10	19 missing observation

V. RESULTS AND DISCUSSIONS

Three (3) data sets were used in this experiment. For wine and iris data, stricken missing data were employed, for breast cancer data, the original missing observations were used as indicated in Table I. With these data sets, the modified imputation method was applied to make the data sets complete. To investigate the effect of the modified imputation method when applied to classification, well known classification algorithms, namely Naïve Bayes, One-R, k-Nearest Neighbor (kNN), C4.5 and Support Vector Machine (SVM) were used. Performance results in terms of precision, receiver operating characteristics (ROC) and accuracy were analyzed and contrasted before and after the modified imputation was applied.

Fig. 4a and 5a show the performance results of each classifier in terms of Precision; Fig. 4b and 5b in terms of ROC, and Fig. 4c and 5c in terms of accuracy. As observed in Fig. 4a, precision increased consistently after imputation was applied in NB, kNN, and SVM for iris data. For wine data, it was kNN which has increased precision for all missing percentage of data after imputation. There may be a slight decrease in some percentage of missing data for other classifiers in both wine and iris data, but differences in the decline is not that remarkable. In terms of ROC, as shown in Fig. 4b, though kNN shows consistent increase after imputation for both iris and wine data, SVM also shows good performance after imputation especially for iris data, and on most percentage of missing data for wine data set. Again, even if there is a slight decrease on other classifier's ROC value after imputation, differences are not remarkable.

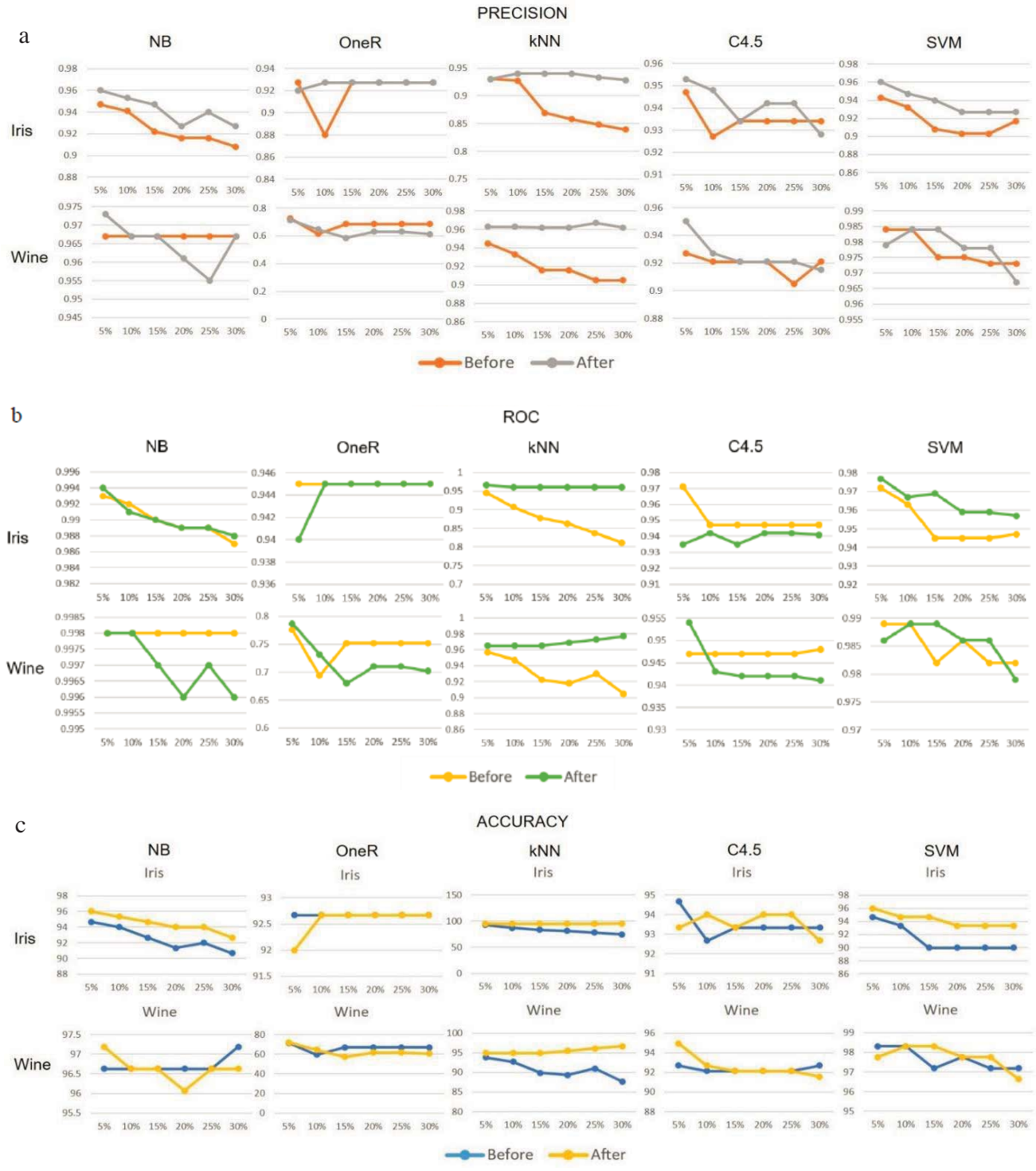

Figure 4. Performance results of classifiers.

Likewise, as shown in Fig. 4c, kNN consistently increased accuracy after imputation for both iris and wine data. But NB and SVM also shows good performance after imputation especially on iris data set. For breast cancer data, evidently, classifier performance increased after imputation. It was also perceived from Fig. 4, for most classifiers, accuracy, precision and ROC values tend to decline as percentage of missing values increases.

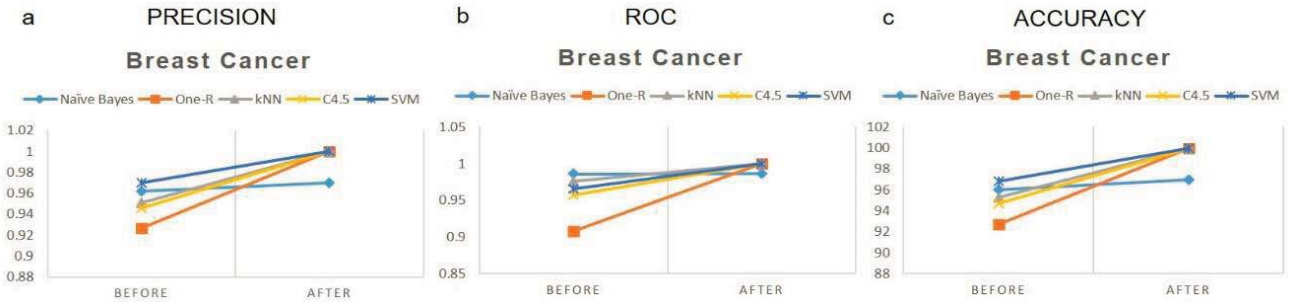

Figure 5. Performance result of breast cancer data.

After the modified imputation was applied, the combined average precision, accuracy and ROC values were computed and results are shown in Table II. Evidently, overall performance increased after imputation using the three performance measures. This generally tells us that when the modified imputation method is applied, it can improve classification performance. Also, based on this, the better performing classifier was determined after the modified imputation was applied and this is illustrated in Fig. 6. SVM has obtained the highest value (0.97628706), followed by kNN (0.97034574), NB (0.96835006), C4.5 (0.95689835), and lastly, One-R (0.86143924), as observed from Table II.

TABLE II. PERFORMANCE SUMMARY

Method	Before Imputation	After Imputation
Naïve Bayes	0.961858556	0.96835006
One-R	0.851432981	0.86143924
kNN	0.910331315	0.97034574
C4.5	0.938803556	0.95689835
SVM	0.959031593	0.97628706

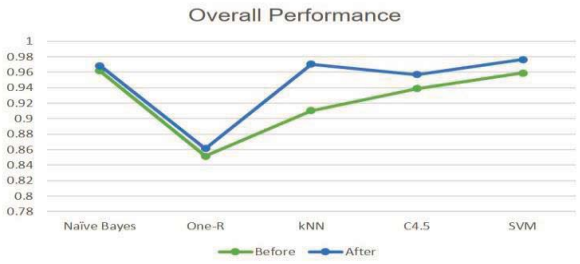

Figure 6. Classifiers' overall performance before and after imputation.

VI. CONCLUSION AND RECOMMENDATIONS

In this study, the modified imputation method was applied to well-known classifiers: Naïve Bayes, One-R, k-Nearest Neighbor (kNN), C4.5, and Support Vector Machine (SVM). Results show that classification performance improves when the modified imputation method is applied to the data sets with missing values during preprocessing. SVM stood out in precision, ROC, and accuracy performance after the modified imputation method was applied. It is noted, however that there are factors that may influence the classifiers' accuracy of prediction namely, the data set, number of instances, class labels, and the like. A classifier that works perfectly in all aspects does not exist. To arrive at the best possible results, it is necessary for the data analyst to compare or combine available techniques.

For future works, it is recommended that other classification performance indicators like F-measure, Mathew's Correlation Coefficient (MCC), and Precision-Recall Curve (PRC) be contrasted before and after the modified imputation is applied. Also, execution time may be considered as one of the performance metrics.

ACKNOWLEDGMENT

Gratefully acknowledge the support and generosity of Bulacan State University, Malolos City, Philippines. Appreciation also goes to the Commission on Higher Education (CHED), K12 Graduate Scholarship Program.

REFERENCES

[1] B. M. Nogueira, T. R. A. Santos, and L. E. Zárate, "Comparison of classifiers efficiency on missing values recovering: Application in a marketing database with massive missing data," *Proc. 2007 IEEE Symp. Comput. Intell. Data Mining, CIDM 2007*, no. Cidm, pp. 66–72, 2007.

[2] S. Kanchana and A. S. Thanamani, "Elevating the Accuracy of Missing Data Imputation Using Bolzano Classifier," vol. 8, no. 1, pp. 138–145, 2016.

[3] J. Sim, J. S. Lee, and O. Kwon, "Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications," vol. 2015, 2015.

[4] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.

[5] S. Ghorbani and M. C. Desmarais, "Performance Comparison of Recent Imputation Methods for Classification Tasks over Binary Data," *Appl. Artif. Intell.*, vol. 31, no. 1, pp. 1–22, 2017.

[6] P. Schmitt, J. Mandel, and M. Guedj, "Biometrics & Biostatistics A Comparison of Six Methods for Missing Data Imputation," vol. 6, no. 1, pp. 1–6, 2015.

[7] P. Liu, L. Lei, and N. Wu, "A Quantitative Study of the Effect of Missing Data in Classifiers," *Comput. Inf. Technol. 2005. CIT 2005. Fifth Int. Conf.*, pp. 28--33, 2005.

[8] L. C. Blomberg and D. D. A. Ruiz, "Evaluating the influence of missing data on classification algorithms in data mining applications," *SBSI 2013 Simp{ó}sio Bras. Sist. Informa{ç}{ã}o*, vol. 0, no. 1, pp. 734–743, 2013.

[9] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," *2013 Fourth Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–7, 2013.

[10] T. N. Phyu, "Survey of Classification Techniques in Data Mining," *Int. MultiConference Eng. Comput. Sci.*, vol. I, pp. 18–20, 2009.

[11] F. M. F. Lobato, V. W. Tadaiesky, I. M. Araújo, and Á. L. de Santana, "An Evolutionary Missing Data Imputation Method for Pattern Classification," *Proc. Companion Publ. 2015 Genet. Evol. Comput. Conf.*, no. July, pp. 1013–1019, 2015.

[12] S. Maddina, "Intelligent Based Imputation Methods for Text Mining Applications to Phishing Attacks," vol. 4, no. 8, pp. 481–485, 2015.

[13] J. Sim, O. Kwon, and K. C. Lee, "Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets," *Expert Syst. Appl.*, vol. 46, no. im, pp. 485–493, 2016.

[14] F. Lobato, C. Sales, I. Araujo, V. Tadaiesky, L. Dias, L. Ramos, and A. Santana, "Multi-objective genetic algorithm for missing data imputation," *Pattern Recognit. Lett.*, vol. 68, pp. 126–131, 2015.

[15] Doreswamy and K. S. Hemanth, "Performance Evaluation of Predictive Engineering Materials Data Sets," *Artif. Intell. Syst. ans Mach. Learn.*, vol. 3, no. 3, pp. 1–8, 2011.

[16] T. R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *Int. J. Comput. Sci. Appl. ISSN 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.

[17] V. Labatut and H. Cherifi, "Evaluation of Performance Measures for Classifiers Comparison," *Ubiquitous Comput. Commun. J.*, vol. 6, pp. 21–34, 2011.

[18] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.," *BMC Genomics*, vol. 13 Suppl 4, no. Suppl 4, p. S2, 2012.

[19] S. Chen, R. Zhan, and J. Zhang, "Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics," *Remote Sens.*, vol. 10, no. 6, 2018.

[20] E. T. Capariño, A. M. Sison, and R. P. Medina, "A Modified Imputation Method to Missing Data as a Preprocessing Technique," *10th IEEE Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag.*, pp. 0–5, 2018.