

Comparing Statistical and Machine Learning Imputation Techniques in Breast Cancer Classification

Imane Chlioui[1], Ibtissam Abnane[1], and Ali Idri[1,2(✉)]

[1] Software Project Management Research Team, ENSIAS, Mohammed V University in Rabat, Rabat, Morocco
imanechlioui@gmail.com, ibtissam_abnane@um5s.net.ma, ali.idri@um5.ac.ma

[2] CSEHS-MSDA, Mohammed VI Polytechnic University, Ben Guerir, Morocco

Abstract. Missing data imputation is an important task when dealing with crucial data that cannot be discarded such as medical data. This study evaluates and compares the impacts of two statistical and two machine learning imputation techniques when classifying breast cancer patients, using several evaluation metrics. Mean, Expectation-Maximization (EM), Support Vector Regression (SVR) and K-Nearest Neighbor (KNN) were applied to impute 18% of missing data missed completely at random in the two Wisconsin datasets. Thereafter, we empirically evaluated these four imputation techniques when using five classifiers: decision tree (C4.5), Case Based Reasoning (CBR), Random Forest (RF), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP). In total, 1380 experiments were conducted and the findings confirmed that classification using imputation based machine learning outperformed classification using statistical imputation. Moreover, our experiment showed that SVR was the best imputation method for breast cancer classification.

Keywords: Missing data imputation · Data mining · Breast cancer

1 Introduction

Breast cancer have been considered as the most dangerous type of cancer, and the second leading cause of death especially among women [1]. It is a tumor that can be malignant or benign, mortal if it is not early diagnosed [2]. The main cause of this disease is not well known, although several factors were identified such as: late age at first child, not breastfeeding, early menarche and late menopause, longer term use of hormone replacement therapy in post-menopausal women and current or recent use of the combined oral contraceptive pill, increased alcohol intake, post-menopausal obesity and reduced physical activity [3, 4]. An early diagnosis can prevent the tumor from spreading to other body parts, thus several clinical methods have been used by practitioners such as: Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), mammography, and Biopsy [1].

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12252, pp. 61–76, 2020.
https://doi.org/10.1007/978-3-030-58811-3_5

Missing data (MD) is an ubiquitous problem in real data, and it can occur due to defective equipment [5], missed entries when filling in data [6], loss of information, human errors [7], patient data safety and policies [8], and information not provided by patients [9]. Several MD techniques were proposed in the literature, and in general imputation proved to be the most efficient solution compared to toleration and deletion [10, 11]. Imputation techniques can be grouped into two groups [12]: i) Statistical imputation techniques common in the literature, many techniques have been proposed including mean [13], hot deck [14] and expectation-maximization [15], ii) since 2000's a new approach has emerged where missing values has been treated as an output of machine learning models [16]. In this framework, the observed data are considered as a training set for the learning model, which is then applied to the data with missing values to impute. K-Nearest Neighbor (KNN) [17], Decision Tree (DT) [18] and Support Vector Regression (SVR) [19] are the most used ML techniques for imputation and achieved great success [20].

With the introduction of ML imputation techniques, several researchers tend to compare the impacts of both statistical and ML techniques on the performance prediction and classification systems. In breast cancer, few papers were published tackling this issue: (1) Acuña and Rodriguez [21] carried out experiments with twelve medical datasets including BC databases, to evaluate the impacts of four MD techniques: deletion, mean imputation, median imputation, and KNN imputation on the misclassification measured in terms of error rate; Their findings suggested that KNN imputation may be the best approach to impute missing values. And (2) Jerez et al. [12] conducted an experiment to evaluate the impacts of several statistical imputation techniques (e.g., mean, hot-deck and multiple imputation), and machine learning imputation techniques (e.g., multi-layer perceptron (MLP), self-organisation maps (SOM) and k-nearest neighbour (KNN)) on an Artificial Neural Network based BC prognosis system. The results showed that according to the Area Under Cover (AUC), the ML imputation techniques outperformed the statistical ones, and Friedman test revealed a significant difference in the observed AUC.

To the best of our knowledge a minority of papers discussed the comparison of imputation techniques for breast cancer classification [22]; moreover none of them used several evaluation metrics to support their findings. This paper investigates whether ML imputation techniques may lead to better results compared to statistical techniques based on different evaluation metrics. Two statistical techniques including mean imputation which is the simplest reference [12] and Expectation Maximization (EM) which assumed to generate unbiased parameter estimates [23]. Two machine learning (ML) imputation techniques, including K-Nearest Neighbor (KNN) which is the most used ML imputation technique [17] and Support Vector Regression (SVR) which have proven to achieve good results for MD imputation [19, 24]. For classification five classifiers were applied over two Wisconsin datasets: Case-based reasoning (CBR), Decision Tree (C4.5), Support Vector Machines (SVM), Random Forest (RF) and Multilayer Perceptron (MLP). And for comparison three evaluation metrics were used: balanced accuracy, Area Under Cover and Kappa value along with Scot-Knott algorithm and Borda count method. In order to draw conclusion on which technique is the best, two research questions were addressed:

RQ1: Does ML imputation techniques significantly outperform statistical imputation ones?

RQ2: Among the ML imputation techniques used, which one achieve the highest performance?

The paper is structured as follows: MD techniques and classification algorithms as well as the datasets used are described in Sect. 2, the experimental design followed is detailed in Sect. 3. The results are presented and discussed in Sect. 4. Finally, Sect. 5 presents conclusion and future work.

2 Materials and Methods

This section describes the techniques used for either handling missing data or classification. As well as the datasets used for the experiments.

2.1 Missing Data Techniques

Thereafter, we briefly present the four imputation techniques used.

- Mean imputation (MI): considered as the simplest method to deal with missing data, widely used as an alternative of Deletion [12]. It consists of replacing the missing data with the mean of all observed cases [25].
- Expectation Maximization (EM): considered as a sophisticated method to deal with missing data. It consists of two steps: the expectation step which uses an initial estimate parameter and conditions it upon the observed variables, the maximization step which provides a new estimation of the parameter using Maximum Likelihood. This process is iterated until convergence [26].
- K-nearest neighbor (KNN): Considered as the most ML technique used for MD imputation [21], based on KNN algorithm. It consists of imputing the missing value considering the K closest instances according to a given distance metric [27].
- Support Vector Regression (SVR): Considered as the regression model of Support Vector Machines (SVM) developed by Vapnik [28]. The SVM is a non-linear algorithm implementing the structural risk minimization inductive principle [29]. They are most known for classification but With the introduction of Vapnik's ε-insensitivity loss function [19], the regression model (SVR) achieved excellent performances [30–32]. However, to the best of our knowledge there is no study investigating the use of SVR imputation for MD in breast cancer datasets.

2.2 Classification Techniques

This study uses five classifiers:

2.2.1 C4.5

C4.5 is a decision tree algorithm developed by Quinlan in 1993 [33]. C4.5 generates classifiers expressed as decision trees; the nodes contain test that will divide-up the training cases, and the results of which permits to decide which bough to follow from

the node. The leaf nodes are the class labels instead of nodes [33]. C4.5 mainly has two parameters [33, 34]:

- CF (confidence factor): affects the confidence with which error rate at the tree nodes is estimated, lower CF values incur heavier pruning.
- MS (minimum numbers of split-off cases): affects the size of the grown tree by disallowing the formation of tree nodes whose number of cases is smaller than the chosen MS parameter; thus, higher MS parameter values lead to grown trees that are of smaller size.

2.2.2 SVM

SVM is a group of supervised learning methods developed by Vapnik in the 90's [34]. It is used to model data not linearly separable. To classify data, the algorithm generates an optimal hyper plane which separates different classes and assigns every input of test set data to one of the defined classes [35, 36]. SVM has three parameters [37]:

- C: is the regularization parameter.
- Kernel: specifies the kernel type to be used in the algorithm. It can be 'linear', 'polynomial', 'Gaussian radial basis function', 'sigmoid', 'precomputed', or a callable.
- Kernel parameter: depends on the chosen kernel. For example: Gamma is the kernel coefficient for 'The Gaussian radial basis function', 'polynomial', and 'sigmoid'.

2.2.3 CBR

Case-based reasoning (CBR) is a method of lazy learning known as a nonparametric method. It consists of aggregating the outputs of the most similar cases in order to predict the output of a new case. It uses as a measure of cases similarity the distance between their data to determine the most similar cases to a given new case. Several distance metrics can be used, such as Hamming distance, Manhattan distance and Euclidean distance [38, 39]. Thereafter, the adaptation step of CBR aggregates the outputs of the k most similar cases in order to predict the output of the new case: the median and the mode are the most used aggregation methods for continuous and categorical outputs respectively [40]. CBR mainly has two parameters some of them are [39]:

- Number of neighbors (K).
- Distance metric to use.

2.2.4 RF

Random forest (RF) algorithm is defined as a generic principle of randomized ensembles of decision trees [41]. It is one of the most powerful techniques in the pattern recognition and machine learning for high-dimensional classification and skewed problems [42]. The accuracy of a forest classifier depends on the strength of each individual tree in the generalized forest and the correlation between them. Indeed, increasing the strength of the individual trees increases the forest accuracy rate while

increasing the correlation reduces the forest accuracy rate [42]. RF has several parameters, important ones are [43]:

- Number of iterations (I).
- Number of features (K).
- Number of execution slots: default = 1
- Seed = 1

2.2.5 MLP

Multi-Layer Perceptron (MLP) is a type of artificial neural network (ANN) that can represent complex input-output relationships [44]. MLP consists of neurons organized in three layers: input, hidden, and output layers which each one performs simple task of information processing by converting received inputs into processed outputs. Although each neuron implements its function slowly and imperfectly, collectively a neural network is able to perform a variety of tasks efficiently and achieve remarkable results [45]. MLP has several parameters, important ones are:

- L: Learning Rate of its back-propagation learning algorithm. (In general within the interval [0, 1], Default = 0.3).
- M: Momentum Rate of its back-propagation algorithm. (In general within the interval [0, 1], Default = 0.2).

2.3 Datasets Description

The experiments were conducted using two datasets collected at the University of Wisconsin–Madison Hospitals [46]. The first one is the Wisconsin breast cancer original dataset, which contains 699 instances periodically collected within the period 1989–1992 [47]; each patient is described by 10 numerical attributes. The second one is the Wisconsin breast cancer prognosis dataset; it contains 198 records and each one represents follow-up data for one breast cancer case, and only includes those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. Each record is described by 35 attributes and 30 among them were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

All cases containing missing data were deleted, which reduced the size of each dataset: 683 instances remain in Wisconsin original and 194 instances in Wisconsin prognosis. Moreover, we normalized the attributes of Wisconsin breast cancer prognosis dataset within the interval [1, 2] in order to avoid bias of attributes' ranges. Note that the attribute values of the Wisconsin breast cancer original dataset were already normalized within the interval [1–10]. Table 1 presents datasets information, including the number of instances and attributes. All the attributes used in this study are numerical.

3 Experimental Design

This section details the process followed for this experiment. As shown in Fig. 1, it consists mainly of four phases: data removal, missing data imputation, generating classifiers and performance evaluation. Each phase is detailed in the following subsections.

Table 1. Datasets description

Database	Instances	Attributes	Attributes type	Source
Wisconsin breast cancer original	194	35	Numeric	[48]
Wisconsin breast cancer prognosis	683	10	Numeric	[48]

3.1 Data Removal

This study aims to evaluate the impact of different imputation methods on the classification of breast cancer datasets. For this purpose, the datasets should be complete to work with, that is why all the missing data already existed were discarded. Thereafter, 18% of MD were randomly induced. The MD are not related to any other value of the existing attributes. Moreover, according to Peng and Lei [25] more than 15% of MD impact severely any interpretation and must be handled using sophisticated techniques, which explain the choice of 18% of MD. It's noteworthy, that the impact of MD percentage on imputation was discussed in a previous article [49]. The findings confirmed that the MD percentage affects negatively the classifier performance.

3.2 Missing Data Imputation

For this step, four imputation techniques were applied to handle the missing values on the datasets resulted from the data removal step.

- Statistical imputation techniques: two statistical techniques were used. The mean imputation which impute using the mean value, and the Expectation-Maximization which apply the Maximum Likelihood estimation method.
- Machine learning imputation techniques: two regression techniques were used. The KNN imputation which is the most famous algorithm used for imputation and SVR imputation which proved to be helpful in effort estimation [19]. Grid search (GS) was used to vary the parameter configuration for each technique according to Table 2.

At the end of this step, we obtained 276 complete datasets: by combining 2 incomplete datasets * (100 complete datasets using SVR (10 variations of C and 10 variations of G) + 36 complete datasets using KNN (4 distance metrics and 9 variations of K) + 1 complete dataset using mean + 1 complete dataset using EM).

3.3 Generating Classifiers

Five classifiers (C4.5, CBR, RF, SVM and MLP) were applied on the generated complete datasets, in order to evaluate and compare the influence of the different imputation techniques used. To fulfill this task, 10-fold cross validation method was used to divide the test and training test. Moreover to guarantee high performance, parameter tuning is used [50]. We opted for GS method to choose the optimal configuration parameter according to Table 2, the best variant with the highest value of the balanced accuracy criterion was retained for the comparison.

At the end of this step, we obtained 1380 classification experiments (276 * 5 = 1380).

3.4 Performance Evaluation

Three different evaluation metrics were considered to evaluate the classifiers and study the convenience of imputing data using both statistical and ML imputation techniques:

– The balanced accuracy rate: it equally weights the value of making accurate predictions in each class in order to avoid bias results caused by imbalanced data [51]. The balanced accuracy rate was evaluated using the Eq. (1).

$$Balanced\ accuracy = \frac{\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}} + \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}}{2} \tag{1}$$

– Kappa value: it represents the pairwise agreement between two different observers (observed accuracy and expected accuracy) [52]. Observed Accuracy is simply the number of instances that were correctly classified, while the expected accuracy is the number of instances of each class along with the number of instances that the classifier classified as true. Its upper limit is +1.00, and its lower limit falls between zero and −1.00 [53]. The kappa statistic value was evaluated using the Eq. (2) [52].

$$K = pA - PE/[1 - PE] \tag{2}$$

$$PA = (TP + TN)/N \tag{3}$$

$$PE = \frac{(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TP} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}{N^2} \tag{4}$$

– Area Under an ROC Cover (AUROC): mostly known as AUC. The Receiver Operating Characteristics (ROC) is a plot of the true positive as a function of the false positive. the AUC is defined as the area under the ROC curve and it is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example [54].

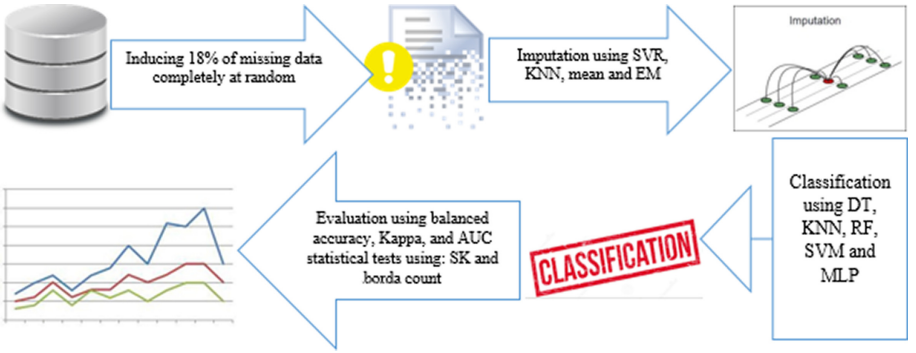

Fig. 1. Experimental design

Table 2. Parameters ranges and optimal configuration of each classifier

MD technique	Parameter ranges	Classifier	Parameter ranges	Optimal configuration for imputation	Optimal configuration for classification
KNN imputation	K = {2 -> 10, increment = 1} Distance metrics: Euclidean, Manhattan, Chebyshev, Minkowsk	C4.5	C = {0.1 -> 1, increment = 0.05}; M = {1 -> 20, increment = 1};	K = 5	C = 0.25 M = 1
		CBR	K = {1 -> 12, increment = 1);	K = 5	K = 9
		RF	I = {10 -> 100, increment = 10); K = {1 -> 10, increment = 1};	K = 5	I = 50 K = 2
		SVM	Kernel = RBFKernel; C = {1 -> 10, increment = 1}; G = {0.01 -> 0.1, increment = 0.01};	K = 7	C = 3 G = 0.01
		MLP	L = {0.1 -> 1, increment = 0.1}; M = {0.1 -> 1, increment = 0.1};	K = 5	M = 0.1 L = 0.1
SVR imputation	Kernel = RBFKernel; C = {1 -> 10, increment = 1}; G = {0.01 -> 0.1, increment = 0.01};	C4.5	C = {0.1 -> 1, increment = 0.05}; M = {1 -> 20, increment = 1};	C = 2 G = 0.01	C = 0.25 M = 1
		CBR	K = {1 -> 12, increment = 1);	C = 5 G = 0.01	K = 9
		RF	I = {10 -> 100, increment = 10); K = {1 -> 10, increment = 1};	C = 2 G = 0.01	I = 80 K = 2
		SVM	Kernel = RBFKernel; C = {1 -> 10, increment = 1}; G = {0.01 -> 0.1, increment = 0.01};	C = 3 G = 0.01	C = 1 G = 0.01
		MLP	L = {0.1 -> 1, increment = 0.1}; M = {0.1 -> 1, increment = 0.1};	C = 2 G = 0.01	M = 0.5 L = 0.1

3.5 Significance Tests

Several evaluation metrics have been used to select the best technique that achieved the highest results. However, different metrics can yield to different conclusions (i.e., if one metric prove that method A is better than method B, another metric can prove that method B is better than method A). Subsequently, two additional methods were used to significantly compare the imputation techniques used:

- Scot-Knott (SK): It's a hierarchical clustering algorithm that permits to distinguish groups with significant F-test [55]. SK is the most frequently used test among those designed for similar purposes [56]. We used SK based on the balanced accuracy rates.
- Borda count method: It's a voting system based on computing the mean rank of each candidates over all voters [57]. We used borda count method based on the three evaluation metrics mentioned above (balanced accuracy, AUC, Kappa).

4 Results and Discussion

This section evaluates and compares the influence of four imputation techniques based on both statistical and ML imputation techniques on the performance of five classifiers, over 18% of MCAR missing data in two Wisconsin breast cancer datasets.

To investigate whether the use of ML imputation techniques (i.e. SVR and KNN) outperform the use of statistical imputation techniques (i.e. mean and EM), Fig. 2, 3 and 4 presents respectively the mean balanced accuracy rates, the mean Kappa and the mean AUC for five classifiers (DT, CBR, SVM, RF and MLP) applied to two breast cancer datasets.

- Based on balanced accuracy rates, ML imputation techniques achieved the highest results compared to statistical imputation techniques for all the five classifiers. SVR and KNN proved to be more efficient for enhancing the classifiers balanced accuracy rates (the mean balanced accuracy rates achieved using RF are for SVR: 88%, for KNN: 84%, for EM: 83% and for mean: 83%). Moreover, SVR outperformed KNN for all the five classifiers (the mean balanced accuracy rates achieved using CBR are for SVR: 83%, for KNN: 81%). While, the results achieved by EM and mean are slightly the same for all the classifiers (the mean balanced accuracy rates achieved using CBR are for EM: 80% and for mean: 80%)
- Based on Kappa rates, ML imputation techniques outperformed once again the statistical imputation techniques for almost all the classifiers, while when using CBR the results were the same for the four imputation techniques (the mean kappa rates achieved using RF are for SVR: 57%, for KNN: 51%, for EM: 45% and for mean: 46%). Furthermore, SVR achieved highest results compared to KNN for C4.5 and RF, while for CBR, MLP and SVM the results were the same (the mean kappa rates achieved using C4.5 are for SVR: 60%, for KNN: 59%). Besides, mean imputation surpassed EM when using RF and SVM, while for C4.5 EM achieved higher results (the mean kappa rates achieved using RF are for EM: 46% and for mean: 45%).

– Based on AUC rates, mean imputation outperformed other techniques when using C4.5, CBR and MLP followed by KNN imputation (the mean AUC rates achieved using C4.5 are for SVR: 78%, for KNN: 77%, for EM: 75% and for mean: 81%). While when using SVM, SVR achieved the highest AUC rate (the mean AUC rates achieved using SVM are for SVR: 81%, for KNN: 80%, for EM: 79% and for mean: 80%). Furthermore, when using RF all the techniques achieved the same AUC rate.

Afterwards, we used SK algorithm to cluster the results based on the balanced accuracy rates. According to Fig. 5, it's observed that the best cluster is the one composed of ML imputation techniques. Next, we applied the borda count voting system based on three evaluation metrics, to rank the techniques belonging to the best SK cluster. From Table 3, it can be seen that SVR obtained the highest number of votes.

To summarize the findings, ML imputation techniques yield to best results compared to statistical imputation techniques. Although implementing ML imputation techniques may be costly, and may need more effort and time, our finding suggest that the performance of classifiers will be improved significantly using ML for MD imputation compared to statistical imputation. Our results align with previous experiments conducted by Perez et al. [12] using ML techniques for MD imputation to improve the prediction accuracy. SVR imputation proved to be more efficient to enhance the performance compared to other techniques [58] not only for breast cancer classification but also in many fields such us software development effort estimation. According to Idri et al. [19] the use of SVR imputation rather than KNN imputation enhance the performance of the prediction performance of both fuzzy and classical analogy-based techniques.

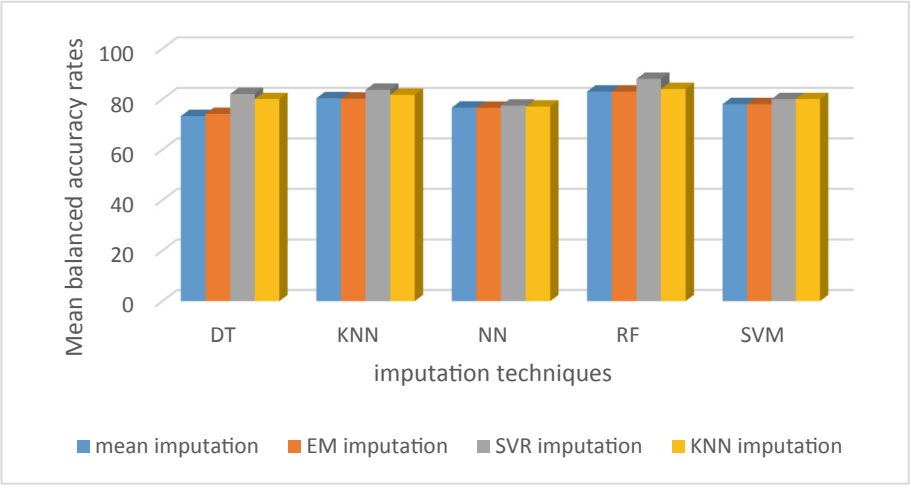

Fig. 2. Mean balanced accuracy rates of C4.5, CBR, RF, MLP and SVM when using mean, EM, KNN and SVR imputation

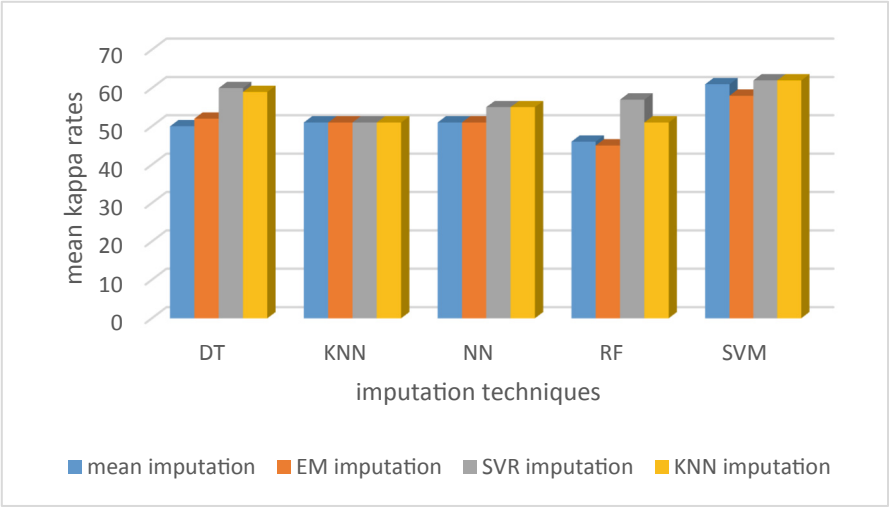

Fig. 3. Mean kappa rates of C4.5, CBR, RF, MLP and SVM when using mean, EM, KNN and SVR imputation

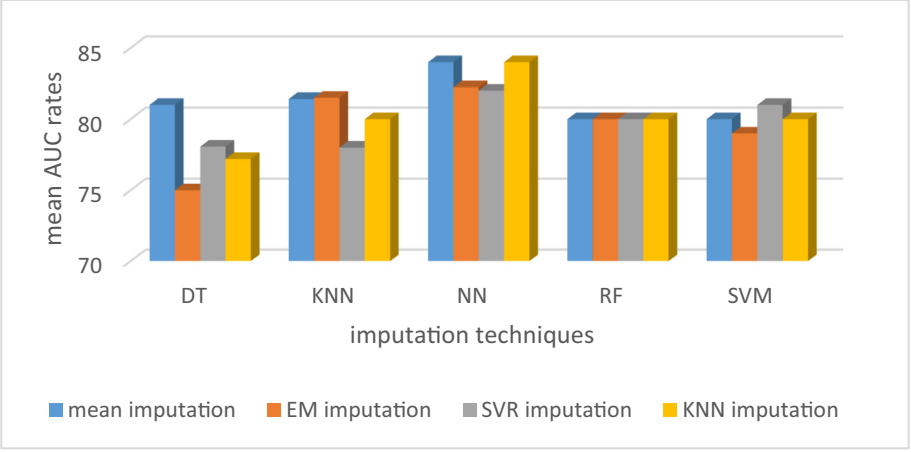

Fig. 4. Mean AUC rates of C4.5, CBR, RF, MLP and SVM when using mean, EM, KNN and SVR imputation

Table 3. Rankings of ML imputation techniques of SK best clusters based on Borda Count

Votes	Rank		
	9	2	4
1	SVR	KNN	SVR/KNN
2	KNN	SVR	

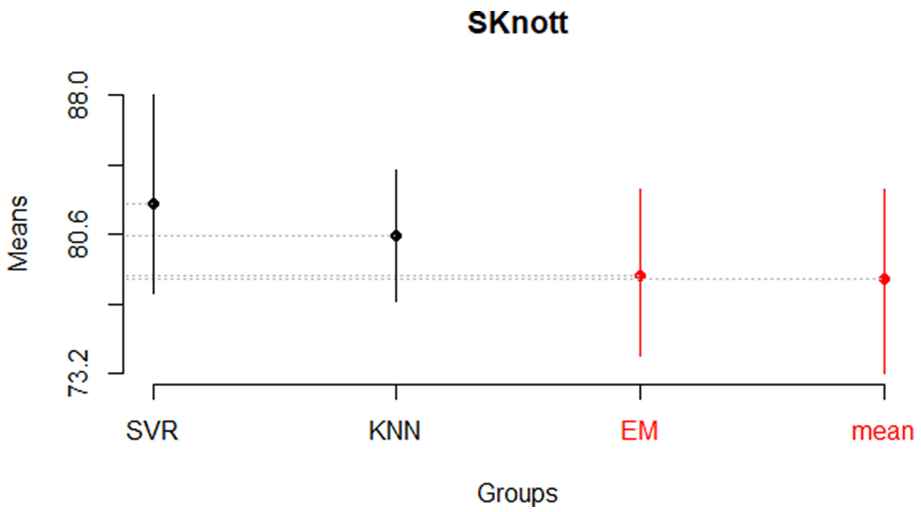

Fig. 5. Plot SK test of imputation techniques used

5 Threats to Validity

This section highlights the internal and external threats of this article.

- Internal validity: Internal validity of this paper are related to the evaluation metrics and method used for the classifiers evaluation. The findings of this study were based on three different evaluation metrics: balanced accuracy, Kappa and AUC. Moreover, we used the SK algorithm to assess the significance of performance differences and Borda count method to rank the best SK cluster techniques. For the evaluation method, 10 cross-validation model was adopted because it is considered as a standard for performance estimation and technique selection [59].
- External validity: The datasets used for this experiment only contain numerical attribute. Further investigations on other datasets are required to discuss categorical attributes. Five classifiers were applied to evaluate the influence of MD imputation techniques in breast cancer classification. Furthermore, the present study only used two statistical (Mean and EM) and two ML (SVR and KNN) imputation techniques, yet other imputation techniques with other classifiers can achieve better results.

6 Conclusion and Future Work

In this study, the impacts of statistical/ML imputation techniques on the five classifiers: C4.5, CBR, RF, SVM and MLP were evaluated over two datasets from the UCI repository: Wisconsin breast cancer original and prognosis datasets. Four imputation techniques: Mean, EM, SVR and KNN were applied on 18% MD.

RQ1: Does ML imputation techniques significantly outperform statistical imputation ones?

Our findings confirm that using ML imputation techniques improved the classifiers performance. However, they are difficult to implement and costly.

RQ2: Among the ML imputation techniques used, which one achieve the highest performance?

The experiments proved that the performance of SVR imputation is slightly superior compared with KNN.

Ongoing research intends to carry out more empirical evaluations of the impact of MD on the performance of breast cancer classification in order to refute or confirm the findings of the present study. Moreover, we intend to investigate other imputation ensembles: homogenous and heterogeneous [60, 61] based on other single imputers such as decision trees [62] and other statistical imputation techniques. Since the present study only deals with numerical attributes, it would be of great interest to deal with missing categorical data too [63].

References

1. Oskouei, R.J., Kor, N.M., Maleki, S.A.: Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges. Am. J. Cancer Res. **7**, 610–627 (2017)
2. Garg, B.: Optimizing number of inputs to classify breast cancer using artificial neural network. J. Comput. Sci. Syst. Biol. **02**, 247–254 (2009). https://doi.org/10.4172/jcsb.1000037
3. Sibbering, M., Courtney, C.A.: Management of breast cancer: basic principles. Surg. (United Kingdom) **34**, 25–31 (2016). https://doi.org/10.1016/j.mpsur.2015.10.005
4. Morimoto, L.M., et al.: Obesity, body size, and risk of postmenopausal breast cancer: the women's health initiative (United States). Cancer Causes Control **13**, 741–751 (2002). https://doi.org/10.1023/A:1020239211145
5. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**, 147–177 (2002). https://doi.org/10.1037/1082-989X.7.2.147
6. Bhat, V.H., Rao, P.G., Krishna, S., Shenoy, P.D., Venugopal, K.R., Patnaik, L.M.: An efficient framework for prediction in healthcare data using soft computing techniques. In: Abraham, A., Mauri, J.L., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) ACC 2011. CCIS, vol. 192, pp. 522–532. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22720-2_55
7. Idri, A., Benhar, H., Fernández-Alemán, J.L., Kadi, I.: A systematic map of medical data preprocessing in knowledge discovery. Comput. Methods Programs Biomed. **162**, 69–85 (2018)
8. Albayrak, M., Turhan, K., Informatics, M., Introduction, I.: A missing data imputation approach using clustering and maximum likelihood estimation. In: IEEE (ed.) 2017 Medical Technologies National Congress (TIPTEKNO), Trabzon, Turkey, pp. 1–4 (2017)
9. Kadi, I., Idri, A., Fernandez-Aleman, J.L.: Knowledge discovery in cardiology: a systematic literature review. Int. J. Med. Inform. **97**, 12–32 (2017). https://doi.org/10.1016/j.ijmedinf.2016.09.005
10. Lang, K.M., Little, T.D.: Principled missing data treatments. Prev. Sci. **19**(3), 284–294 (2016). https://doi.org/10.1007/s11121-016-0644-5
11. Idri, A., Abnane, I., Abran, A.: Missing data techniques in analogy-based software development effort estimation. J. Syst. Softw. **117**, 595–611 (2016). https://doi.org/10.1016/j.jss.2016.04.058

12. Jerez, J.M., et al.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. **50**, 105–115 (2010). https://doi.org/10.1016/j.artmed.2010.05.002
13. Gayathri, B.M., Sumathi, C.P.: Mamdani fuzzy inference system for breast cancer risk detection. In: 2015 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015, Madurai, Tamilnadu, India, pp. 1–6 (2016)
14. Myers, T.A.: Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. Commun. Methods Meas. **5**, 297–310 (2011). https://doi.org/10.1080/19312458.2011.624490
15. Barzi, F., Woodward, M.: Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am. J. Epidemiol. **160**, 34–45 (2004). https://doi.org/10.1093/aje/kwh175
16. Liu, Y., Gopalakrishnan, V.: An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. Data (2017). https://doi.org/10.3390/data2010008
17. Penone, C., et al.: Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol. Evol. **5**, 961–970 (2014). https://doi.org/10.1111/2041-210X.12232
18. Vateekul, P., Sarinnapakorn, K.: Tree-based approach to missing data imputation. In: ICDM Workshops 2009 - IEEE International Conference on Data Mining, pp. 70–75 (2009)
19. Idri, A., Abnane, I., Abran, A.: Support vector regression-based imputation in analogy-based software development effort estimation. J. Softw. Evol. Process. **30**, 1–23 (2018). https://doi.org/10.1002/smr.2114
20. Wu, X., Akbarzadeh Khorshidi, H., Aickelin, U., Edib, Z., Peate, M.: Imputation techniques on missing values in breast cancer treatment and fertility data. Health Inf. Sci. Syst. **7**(1), 1–8 (2019). https://doi.org/10.1007/s13755-019-0082-4
21. Acuña, E., Rodriguez, C.: The treatment of missing values and its effect on classifier accuracy. In: Banks, D., McMorris, F.R., Arabie, P., Gaul, W. (eds.) Classification, Clustering, and Data Mining Applications. STUDIES CLASS, pp. 639–647. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-642-17103-1_60
22. Chlioui, I., Idri, A., Abnane, I.: Data preprocessing in knowledge discovery in breast cancer: systematic mapping study. Comput. Methods Biomech. Biomed. Eng. Imaging Vis., 1–15
23. Musil, C.M., Warner, C.B., Yobas, P.K., Jones, S.L.: A comparison of imputation techniques for handling Missing data. West. J. Nurs. Res. **24**, 815–829 (2002). https://doi.org/10.1177/019394502762477004
24. Santos, M.S., Soares, J.P., Henriques Abreu, P., Araújo, H., Santos, J.: Influence of data distribution in missing data imputation. In: ten Teije, A., Popow, C., Holmes, J.H., Sacchi, L. (eds.) AIME 2017. LNCS (LNAI), vol. 10259, pp. 285–294. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59758-4_33
25. Peng, L., Lei, L.: A review of missing data treatment methods. Int. J. Intell. Inf. Manag. Syst. Technol. **1**, 412–419 (2005)
26. Moon, T.K.: The expectation-maximization algorithm (1996)
27. Chlioui, I., Idri, A., Abnane, I., de Gea, J.M.C., Fernández-Alemán, J.L.: Breast cancer classification with missing data imputation. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST 2019. AISC, vol. 932, pp. 13–23. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16187-3_2
28. Vapnik, V.: The Nature of Statistical Learning Theory (2013)
29. Debasish, B., Srimanta, P., Dipak Chandra, P.: Support vector regression. Neural Inf. Process. Lett. Rev. **11**, 699–708 (2007)

30. Drucker, H., Surges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: Advances in Neural Information Processing Systems (1997)
31. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression (2004)
32. Müller, K.-R., Smola, A.J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., Vapnik, V.: Predicting time series with support vector machines. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 999–1004. Springer, Heidelberg (1997). https://doi.org/10.1007/BFb0020283
33. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, Amsterdam (1992)
34. Witten, I.H., Frank, E., Hall, M.A.: Data Mining. Elsevier, Amsterdam (2011)
35. Alpaydın, E.: Introduction to machine learning, London (2014)
36. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel Based Learning Methods. Cambridge University Press, Cambridge (2000)
37. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**, 121–167 (1998). https://doi.org/10.1023/A:1009715923555
38. Marsilin, J.R.: An efficient CBIR approach for diagnosing the stages of breast cancer using KNN classifier. Bonfring Int. J. Adv. Image Process. **2**, 01–05 (2012). https://doi.org/10.9756/bijaip.1127
39. Odajima, K., Pawlovsky, A.P.: A detailed description of the use of the kNN method for breast cancer diagnosis. In: 2014 7th International Conference on BioMedical Engineering and Informatics, BMEI 2014, Dalian, China, pp. 688–692 (2014)
40. Kowarik, A., Templ, M.: Imputation with the R Package VIM. J. Stat. Softw. **74**, 1–16 (2016). https://doi.org/10.18637/jss.v074.i07
41. Hu, S., Liang, Y., Ma, L., He, Y.: MSMOTE: improving classification performance when training data is imbalanced. In: 2nd International Workshop on Computer Science and Engineering, WCSE 2009, Qingdao, China, pp. 13–17 (2009)
42. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
43. Pavlov, Y.L.: Random forest. In: Probabilistic Methods in Discrete Mathematics, pp. 11–18 (2000)
44. Ghosh, S., Mondal, S., Ghosh, B.: A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In: 1st International Conference on Automation, Control, Energy and Systems, ACES 2014, Hooghly, West Bengal, India, pp. 1–4 (2014)
45. Brockmann, D., Hufnagel, L., Geisel, T.: Data Mining and Knowledge Discovery Handbook. Springer, Boston (2006)
46. Jhajharia, S., Varshney, H.K., Verma, S., Kumar, R.: A neural network based breast cancer prognosis model with PCA processed features. In: 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, Jaipur, India, pp. 1896–1901 (2016)
47. Song, Q., Shepperd, M., Chen, X., Liu, J.: Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. J. Syst. Softw. **81**, 2361–2370 (2008). https://doi.org/10.1016/j.jss.2008.05.008
48. Dua, D., Graff, C.: UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
49. Chlioui, I., Idri, A., Abnane, I., de Gea, J.M.C., Fernández-Alemán, J.L.: Breast cancer classification with missing data imputation (2019)
50. Idri, A., Hosni, M., Abnane, I., Carrillo de Gea, J.M., Fernández Alemán, J.L.: Impact of parameter tuning on machine learning based breast cancer classification. In: World Conference on Information Systems and Technologies, Galicia, Spain, pp. 115–125 (2019)

51. García, V., Mollineda, R.A., Sánchez, J.S.: Index of balanced accuracy: a performance measure for skewed class distributions. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 441–448. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02172-5_57
52. Jonsdottir, T., Hvannberg, E.T., Sigurdsson, H., Sigurdsson, S.: The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. Expert Syst. Appl. **34**, 108–118 (2008). https://doi.org/10.1016/j.eswa.2006.08.029
53. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960). https://doi.org/10.1177/001316446002000104
54. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Advances in Neural Information Processing Systems, pp. 313–320 (2004)
55. Jelihovschi, E., Faria, J.C., Allaman, I.B.: ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. TEMA (São Carlos) **15**, 003 (2014). https://doi.org/10.5540/tema.2014.015.01.0003
56. Hosni, M., Idri, A., Abran, A., Nassif, A.B.: On the value of parameter tuning in heterogeneous ensembles effort estimation. Soft. Comput. **22**(18), 5977–6010 (2017). https://doi.org/10.1007/s00500-017-2945-4
57. Vuurpijl, L., Schomaker, L.: An overview and comparison of voting methods for pattern recognition. In: IEEE (ed.) Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 195–200 (2002)
58. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**, 263–282 (2010). https://doi.org/10.1007/s00521-009-0295-6
59. Salzberg, S.L.: On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Discov. **1**, 317–328 (1997). https://doi.org/10.1023/A:1009752403260
60. Hosni, M., Abnane, I., Idri, A., de Gea, J.M.C., Alemán, J.L.F.: Reviewing ensemble classification methods in breast cancer. Comput. Methods Programs Biomed. **177**, 89–112 (2019). https://doi.org/10.1016/j.cmpb.2019.05.019
61. Abnane, I., Hosni, M., Idri, A., Abran, A.: Analogy software effort estimation using ensemble KNN imputation. In: IEEE (ed.) 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Kallithea, Greece, pp. 228–235 (2019)
62. Twala, B.: An empirical comparison of techniques for handling incomplete data using decision trees. Appl. Artif. Intell. **23**, 373–405 (2009). https://doi.org/10.1080/08839510902872223
63. Abnane, I., Idri, A.: Improved analogy-based effort estimation with incomplete mixed data. In: Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, Poznań, Poland, pp. 1015–1024 (2018)