JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 1039-1051 (2021)
DOI: 10.6688/JISE.202109_37(5).0004

Ensemble Case based Reasoning Imputation in Breast Cancer Classification*

IMANE CHLIOUI[1], ALI IDRI[1,2], IBTISSAM ABNANE AND MAHMOUD EZZAT[2]
[1]*Software Project Management Research Team*
ENSIAS, Mohammed V University in Rabat
Rabat, 10112 Morocco
[2]*MSDA, Mohammed VI Polytechnic University*
Ben Guerir, 43150 Morocco
E-mail: imanechlioui@gmail.com; ali.idri@um5.ac.ma;
ibtissam_abnane@um5s.net.ma; mahmoud.ezzat@um6p.ma

Missing Data (MD) is a common drawback that affects breast cancer classification. Thus, handling missing data is primordial before building any breast cancer classifier. This paper presents the impact of using ensemble Case-Based Reasoning (CBR) imputation on breast cancer classification. Thereafter, we evaluated the influence of CBR using parameter tuning and ensemble CBR (E-CBR) with three missingness mechanisms (MCAR: missing completely at random, MAR: missing at random and NMAR: not missing at random) and nine percentages (10% to 90%) on the accuracy rates of five classifiers: Decision trees, Random forest, *K*-nearest neighbor, Support vector machine and Multi-layer perceptron over two Wisconsin breast cancer datasets. All experiments were implemented using Weka JAVA API code 3.8; SPSS v20 was used for statistical tests. The findings confirmed that E-CBR yields to better results compared to CBR for the five classifiers. The MD percentage affects negatively the classifier performance: as the MD percentage increases, the accuracy rates of the classifier decrease regardless the MD mechanism and technique. RF with E-CBR outperformed all the other combinations (MD technique, classifier) with 89.72% for MCAR, 87.08% for MAR and 86.84% for NMAR.

Keywords: breast cancer, ensemble, CBR imputation, missing data, classification

1. INTRODUCTION

Breast Cancer (BC) is a major public health challenge worldwide, and the second leading cause of death among women [1]. Breast Cancer occurs when abnormal cells in breast tissue form a tumor by mutation in the Deoxyribo Nucleic Acid (DNA) [2]. Those tumors can be benign or malignant. Benign tumors are similar to normal in appearance, they grow slowly and do not present any harm to the health. Malignant tumors are cancerous and can spread beyond the original tumor to other parts of the body [3].

Nowadays, the emergence of Data mining (DM) has helped to assist doctors in several subfields of medicine such as cardiology [4], endocrinology [5], and oncology [6]. Specifically, DM techniques have been actively used to assist doctors in the process of BC diagnosis and treatment. The mapping study of Ezzat and Idri [7] on the use of data analytics techniques in BC treatment found that classification is the most frequent DM objective discussed. Among the classification techniques used, Decision Tree, fuzzy methods and Support Vector Machine (SVM) are the techniques that gained more attention during the

Received August 31, 2020; revised November 16, 2020; accepted December 22, 2020.
Communicated by Maria José Sousa.
* 8th World Conference on Information Systems and Technologies, Montenegro, Sponsors: IEEE SMC, AISTI, GIIM, ITMA, UAc UNIVERSIDADE DOS Açores.

years [8]. Vrigazova [9] proposed ANOVA-BOOTSTRAP-RBF-SVM an improvement of SVM to enhance the quality of BC diagnosis, using RBF Kernel and a grid of integer values. The best accuracy (99.6%) was achieved with C = 32. Fatih [10] established an adequate model by revealing the predictive factors of early-stage breast cancer patients, using an ensemble of statistical visualization techniques to understand the correlation between features and select the most relevant ones. Thereafter, classification techniques were applied, and Logistic regression achieved the highest accuracy result up to 98.1%, compared to the other classifiers.

Usually, classification models for BC diagnosis are built using medical data collected from hospitals. However, medical data often contain MD which is a major limitation when applying classification techniques in BC diagnosis. It can distort the analysis by introducing a bias into the classification process and affect the patient survival rate [11]. Different techniques have been developed to deal with MD ranging from deletion to imputation. Case Base Reasoning (CBR) is the most used imputation technique [12]. However, none of the imputation techniques has proved to be the best under all circumstances, since their performances depend on the dataset and the classifier used [11, 13]. Thus, it would be more fruitful to combine multiple imputers instead of using a single one. A combination of more than one single predictor into an ensemble under a specific combination rule is called ensemble prediction. To the best of our knowledge, few studies investigated the use of ensembles to impute medical MD and none has yet focused on BC classification [5, 14], which motivates this research. Within this context, the present study investigates the use of ensemble based CBR imputation for BC diagnosis. This consists of constructing and evaluating a CBR ensemble based imputation (E-CBR), whose members are different variants of a single CBR imputation (*i.e.* different parameter settings), when used with five classifiers: K-Nearest Neighbor (KNN), random forest (RF), Decision tree C4.5, SVM and multi-layer perceptron (MLP). Moreover, we investigate whether the optimization of CBR parameters using a grid search during imputation is helpful in BC classification. Thereafter, we compare the performances, measured in terms of balanced accuracy, of the five classifiers using E-CBR with those using single CBR imputation. The empirical evaluations of CBR and E-CBR used three MD missingness mechanisms (MCAR, MAR, NMAR), nine MD percentages (from 10% to 90%), and two datasets: Wisconsin breast cancer original and Wisconsin breast cancer prognosis. They were performed using the experimental process proposed by Idri *et al.* [15]. Two research questions were addressed:

RQ1: Compared to single CBR, does E-CBR positively affect the accuracy of the five classifiers? Is the improvement of the performance of the five classifiers provided by E-CBR over single CBR significant?
RQ2: Is there any combinations of MD techniques and classifiers which perform better than other? Do these combinations depend on the type of MD?

The main contributions of this study are as follow: (1) Evaluating the use of parameter tuning for CBR imputation in BC classification; (2) Proposing and evaluating ensemble based CBR imputation for BC classification; (3) Comparing the influence of single CBR and E-CBR on the performance of BC classification.

This paper is structured as follows. Section 2 details the experimental design followed in this study as well as the datasets used. Section 3 presents the results and discusses the findings. Section 4 concludes the paper and suggests further research lines.

2. MATERIAL AND METHODS

This section describes the datasets used, the performance criteria to evaluate each classifier, and the process to carry out the different experiments.

2.1 Datasets Description

The experiments were conducted using the datasets collected at the University of Wisconsin-Madison Hospitals [16]. The first one is the Wisconsin breast cancer original dataset. The second one is the Wisconsin breast cancer prognosis dataset. All cases containing missing data were deleted, which reduced the size of each dataset: 683 instances 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.

2.2 Experimental Design

Fig. 1 shows the experimental design followed in the present study. This process consists of four main phases: data removal, complete dataset generation, generating classifiers, performance evaluation and statistical tests. This study evaluates the performance of five classifiers with CBR and E-CBR, nine percentages of MD (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%), and three different missingness mechanisms (MCAR, MAR and NMAR) over two datasets. Each step of this process is detailed in the following subsections.

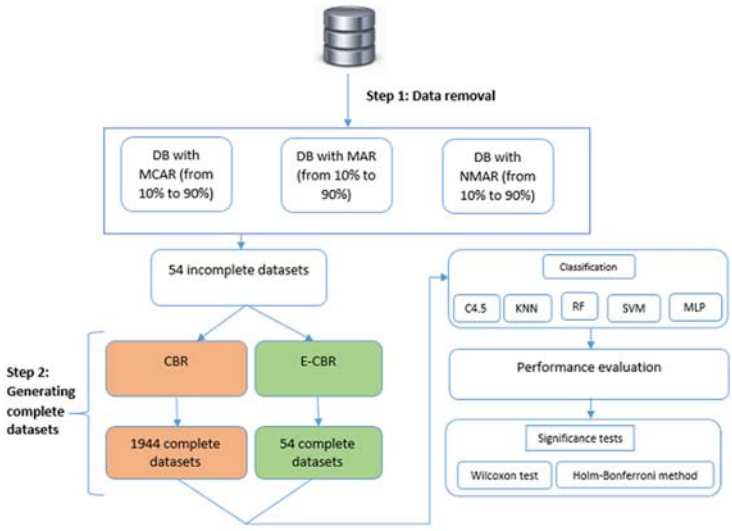

Fig. 1. Experimental design.

(A) Data Removal

We firstly removed all MD already existing in the datasets. Thereafter, the MD were artificially generated using the three missingness mechanisms: (1) MCAR: The MD was

induced completely at random for each variable; (2) MAR: A causative attribute was randomly selected for each dataset: cell_shape_uniformity and lymph_node_status for Wisconsin original and Wisconsin prognosis respectively. First, the instances were sorted in an ascending order of the causative attribute. Thereafter, the datasets were split into three equal subsets, and the MD were generated as follows: (a) 60% * *p* assigned randomly to the first subset; (b) 40% * *p* assigned to the second subset; (3) NMAR: This mechanism is similar to MAR, but instead of inducing MD to all attributes depending on the causative variables, only the causative variable loses values. The causative variables are the same selected for the MAR mechanism. For each missingness mechanism, 9 percentages (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%) were used, which gave us a total of 54 incomplete datasets (54 = 3 MD mechanisms * 9 percentages * 2 datasets).

(B) Complete Dataset Generation

In this step, two MD techniques (E-CBR, CBR) were applied to generate complete datasets by imputing missing values of the incomplete data sets of Step 1.

CBR: Grid search consists of tuning every parameter of a classifier over a predefined range and then select the configuration that provides the best performance of the classifier [17]. Thus, for each incomplete dataset, we used GS to set different parameter of CBR imputation to generate different complete datasets. This study used a GS on two parameters of CBR: distances metric and number of analogies (*k*). Distance metrics can be: Euclidean, Manhattan, Minkowski and Chebyshev while the number of analogies was varied from 2 to 10 with increment 1. This leads to obtain 36 complete datasets for each incomplete dataset.

Ensemble based CBR imputation (E-CBR): For each dataset, E-CBR combines 36 single CBR variants using four distance metrics and 9 values of *k* (36 = 4 (distances) ∗ 9 (values of *k*)). Thereafter, E-CBR uses the median of these 36 imputed values to generate the final E-CBR imputed value. This leads to obtain one complete dataset for each incomplete dataset. Note that each single CBR variant (*i.e.* one single CBR configuration) generated one complete dataset, which in total gave 36 complete datasets. Whereas using ensemble CBR, generated only one complete dataset. This implies that the time consumed to classify one complete dataset generated by an ensemble CBR is 36 times less than classifying all (*i.e.* 36) the complete datasets generated by single CBR with grid search.

At the end of this step, we obtained 1998 complete datasets (54 + 36 (possible configuration) ∗ 54 = 1998).

(C) Generating Classifiers

Five classifiers (C4.5, CBR, RF, SVM and MLP) were used in order to evaluate the influence of the two MD techniques CBR and E-CBR on the classification accuracy. The 10-fold cross validation was used for the evaluation process. This cross-validation method aims to split up the dataset into 10 equal subsets, in which one fold is used as the test set and the remaining folds are used as training sets. The process is repeated ten times, and it permits to every observation point gets to be in a test set exactly once, and gets to be in a training set nine times [18]. After applying the five classifiers to the 1998 complete data sets, we obtained 9990 classification experiments (9990 = 1998*5). For each classifier, the grid search method was used to vary the classifiers parameters according to Table 1. The

best variant of each classifier (*i.e.* with the highest value of the balanced accuracy criterion) was retained for the classifiers comparison.

Table 1. The best variant of each classifier (*i.e.* with the highest value of the balanced accuracy criterion) was retained for the classifiers comparison.

Algorithm	Parameters ranges	Optimal configuration
C4.5	C={0.1->5, increment=0.1}; M={10->100, increment=10};	C=0.25, M=2
KNN	K={1->12, increment =1}	K=1
RF	I={100->1000, increment=100}; {K =1->5, increment=1}; Seed = 1	I=100, K=1 Seed=1
SVM	Kernel=RBFKernel; C={100->200, increment =10}; G={0.01->0.1, increment=0.01}	Kernel=RBFKernel C=1, G=0.01
MLP	L= {0.1>1, increment=0.1}M={0.1->1, increment=0.1}	L=0.3, M=0.2

(D) Performance Evaluation

To evaluate the performance of the five classifiers, the balanced accuracy measure was used; it represents the average of sensitivity and specificity. This equally weights the value of making accurate predictions in each class. Note that the balanced accuracy rate criterion was used, in order to avoid biased results caused by imbalanced data [19].

(E) Significance Test

In order to compare the classifiers performance using CBR and E-CBR, the Wilcoxon statistical tests were performed. It is a non-parametric test that compares two paired samples and analyzes the differences between each set of pairs to assess either there is a significant difference or not [20]. The statistical tests were performed at $\alpha = 0.05$ significance level. The Holm-Bonferroni method was adopted to support the Wilcoxon test results [21].

3. RESULTS AND DISCUSSION

This section presents and discusses the results of the empirical evaluations carried out in this research. First, we report and discuss the performances of the five classifiers when using CBR and E-CBR (RQ1). Moreover, we evaluate whether the performances of classifiers are statistically significant using hypothesis testing. Next, we discuss if there is suitable combinations of classifiers and MD techniques to deal with MD in BC. All the empirical evaluations were coded using the WEKA (3.8.0) API code [22].

4.1 RQ1: Compared to single CBR, does E-CBR positively affect the accuracy of the five classifiers? Is the improvement of the performance of the five classifiers provided by E-CBR over single CBR significant?

This section assesses and compares the impacts of CBR and E-CBR on the balanced accuracy rates of the five classifiers. Figs. 1-2 (a)-(c) and 4-5 (a)-(c), and Figs. 3 (a)-(c) present the mean balanced accuracy rates of C4.5, KNN, RF, SVM, and MLP respectively over to the two datasets using CBR and E-CBR, three MD mechanisms, and nine MD

percentages. Moreover, Wilcoxon statistical tests were performed in order to assess whether the balanced accuracy rates of classifiers are significantly influenced by the MD mechanisms and techniques. To do that, the following hypothesis was drawn: The balanced accuracy of each classifier is not affected by using E-CBR rather than CBR.

For the MCAR mechanism, Figs. 2-6 (a) show that the classification accuracy of all classifiers highly improved when using E-CBR compared to CBR regardless the MD percentage (In average, the mean balanced accuracy rates were for C4.5: 81.23% with E-CBR and 77.63% with CBR, for KNN 82.59% with E-CBR and 81.20% with CBR, for RF: 89.72% with E-CBR and 79.72% with CBR, for SVM: 80.74% with E-CBR and 79.35% with CBR, and for MLP: 78.85% with E-CBR and 73.59%% with CBR). Moreover, we noticed that the balanced accuracy of each classifier with either CBR or E-CBR decreased as the MD percentage increased (mean balanced accuracy rates at 10% and 90% were for C4.5: 84.06% and 72.69% respectively with E-CBR, and 82.29% and 71.62% respectively with CBR; for KNN: 87.10% and 81.55% respectively with E-CBR, and 84.35% and 80.08% respectively with CBR; for RF: 90.80% and 89.32% respectively with E-CBR, and 81.56% and 79% respectively with CBR; for SVM: 82.93% and 79.22% respectively with E-CBR, and 82.4% and 77.33% respectively with CBR; and for MLP: 80.84% and 76.57% respectively with E-CBR, and 73.69% and 73.5% respectively with CBR). According to the Wilcoxon test, the classification results achieved when using E-CBR are significantly better than the results achieved when using CBR ($p(\alpha) = 0.008$ for C4.5; $p(\alpha) = 0.008$ for KNN; $p(\alpha) = 0.008$ for RF; $p(\alpha) = 0.008$ for SVM; and $p(\alpha) = 0.008$ for MLP).

For the NMAR mechanism, we observe from Figs. 2-6 (c), that the use of E-CBR yield to better balanced accuracy rates compared to the use of CBR whatever the classifier and the MD percentage, (In average, the mean balanced accuracy rates were: for C4.5: 78.01% with E-CBR and 74.68% with CBR, for KNN: 80.50% with E-CBR and 75.39% with CBR, for RF: 86.84% with E-CBR and 81.20% with CBR, for SVM: 79.35% with E-CBR and 77.95% with CBR, and for MLP: 80% with E-CBR and 79.17% with CBR). In contrast with the previous mechanisms the balanced accuracy rates of each classifier using either E-CBR or CBR was constant for almost all the MD percentages with a slight minimization between the first percentages for some classifiers (the mean balanced accuracy rates at 10% and 90% were: for C4.5: 78.01% and 78.01% respectively with E-CBR, and 74.72% and 74.67% respectively with CBR; for KNN: 80.5% and 80.5% respectively with E-CBR, and 75.39% and 75.39% respectively with CBR; for RF: 86.84% and 86.84% respectively with E-CBR, and 79.04% and 77.76% respectively with CBR, For SVM: 79.35% and 79.35% respectively with E-CBR, and 77.95% and 77.95% respectively with CBR; for MLP: 80.57% and 79.92% respectively with E-CBR, and 79.17% and 79.17% respectively with CBR). The Wilcoxon test confirms that the classification results when using E-CBR has significantly outperformed the results when using CBR for all the classifiers ($p(\alpha) = 0.006$ for C4.5; $p(\alpha) = 0.007$ for KNN; $p(\alpha) = 0.007$ for RF; $p(\alpha) = 0.003$ for SVM; and $p(\alpha) = 0.004$ for MLP).

Under MCAR and MAR, the MD percentage impacts negatively the balanced accuracy rates achieved by all the five classifiers, though the classifiers maintain acceptable rates regardless the MD techniques except C4.5 with CBR (at 90% of MD the balanced accuracy rates achieved by RF, KNN, SVM, C4.5 and MLP under MCAR are respectively: E-CBR: 89.32%, 81.55%, 79.22%, 79.62% and 73.5%, CBR: 79%, 80.08%, 77.33%, 71.62% and 76.57%).

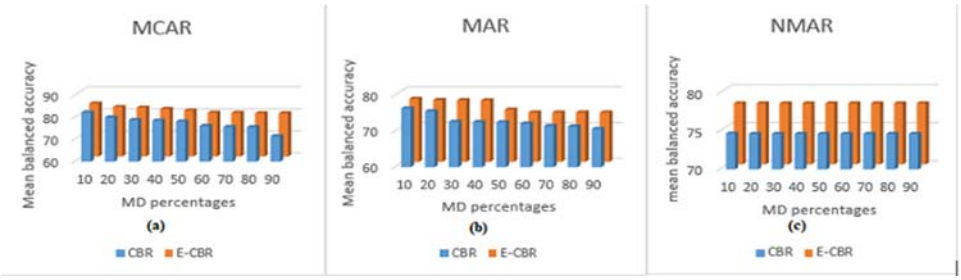

Fig. 2. Mean balanced accuracy rates of C4.5 using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

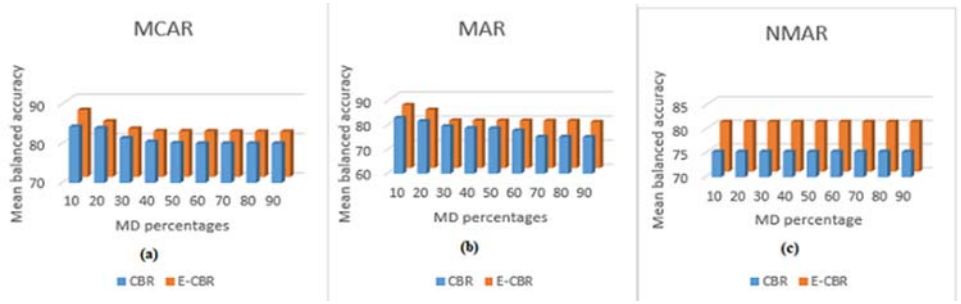

Fig. 3. Mean balanced accuracy rates of KNN using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

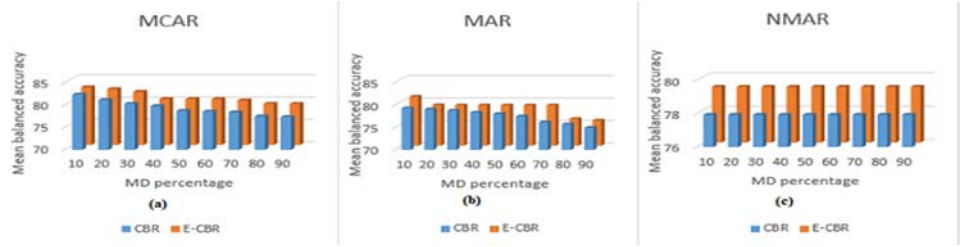

Fig. 4. Mean balanced accuracy rates of SVM using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

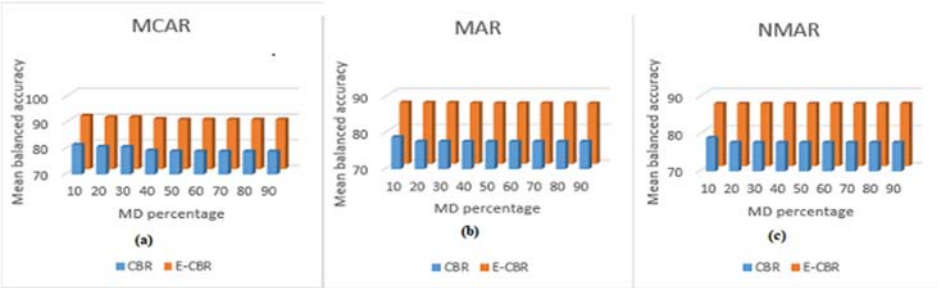

Fig. 5. Mean balanced accuracy rates of RF using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

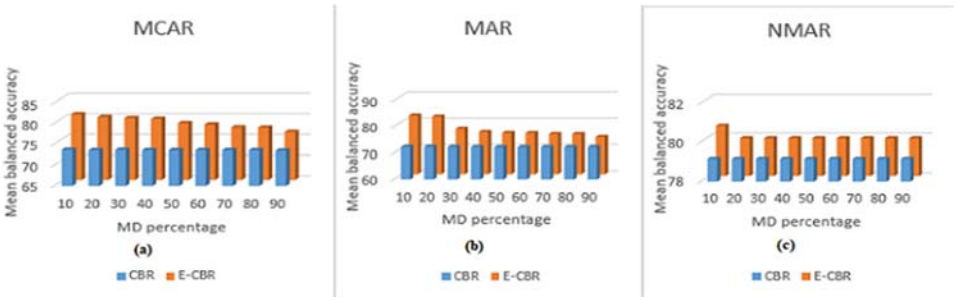

Fig. 6. Mean balanced accuracy rates of MLP using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

To sum up the findings of RQ1, we conclude that:

1. The MD percentage influences negatively the balanced accuracy rates of the classifiers. As long as the MD percentage increases the balanced accuracy rate decreases; which could be due to the fact that the more the dataset contains MD, the more the classification results are biased. For instance, imputing 10% of MD is more reliable due to the remaining large sample of instances, unlike imputing 90% of MD that can bias the dataset [23].
2. The use of E-CBR enhanced the balanced accuracy rate of the five classifiers compared to CBR, regardless the missingness mechanism and the MD percentage. This can be explained by the fact that using ensemble imputation permits to use the power of each single imputer, and employs it to enhance the overall performance of the proposed method [24].
3. In general, MCAR presents the highest balanced accuracy rates. This mechanism is influenced by the randomness, since the MD are induced randomly. Moreover, NMAR presented slightly better results than MAR which is due to the fact that MD are related to the observed variables.

3.2 RQ2: Is there any combinations of MD techniques and classifiers which perform better than other? Do these combinations depend on the type of MD?

This section compares the mean balanced accuracy rates of the five classifiers C4.5, KNN, RF, SVM and MLP, using CBR/E-CBR, three MD mechanisms, and nine MD percentages. The aim is to investigate if there is a suitable combination of classifier/MD technique/MD mechanism to use for breast cancer classification. Fig. 7 shows the mean balanced accuracy values of each classifier using CBR/E-CBR, three MD mechanisms. Thereafter, we used the Wilcoxon and Holm-Bonferroni statistical t-tests in order to evaluate the significance of the differences in the balanced accuracy results:

The combination of RF with E-CBR achieved the highest mean balanced accuracy rates regardless the MD mechanism and MD percentage (In average, the balanced accuracy rates obtained with E-CBR under MCAR are: for RF 89.72%, for KNN 82.59%, for C4.5 81.23%, for SVM 80.73% and for MLP 78.85%). Wilcoxon and Bonferroni tests confirm that RF significantly outperformed the other classifiers (for RF with E-CBR under MCAR, C4.5 with E-CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0166$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.025$; SVM with E-CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0125$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0166$; KNN with

E-CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0125$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0166$; and MLP with E-CBR: $p(\alpha) = 0.008/p(\alpha') = 0.025$, CBR: $p(\alpha) = 0.007/\ p(\alpha') = 0.0125$). Moreover, RF combined with CBR outperformed MLP using E-CBR under MCAR, and C4.5 under MAR regardless MD percentage (In average, the mean balanced accuracy rates obtained: under MCAR using RF with CBR and MLP with E-CBR were respectively: 79.72%, 78.85%; under MAR using RF with CBR and C4.5 with E-CBR were respectively: 77.80%, 75.49%). The differences were statistically significant (for MLP with E-CBR under MCAR: $p(\alpha) = 0.015/p(\alpha') = 0.05$; for C4.5 with E-CBR under MAR: $p(\alpha) = 0.008/p(\alpha') = 0.025$).

The combination KNN with E-CBR comes next and outperformed C4.5, SVM and MLP regardless the MD percentage and mechanism (under MCAR the balanced accuracy rates obtained in average are: for KNN with E-CBR: 82.59%, for C4.5: 81.23% with E-CBR and 77.63% with CBR, for SVM: 80.73% with E-CBR and 79.35% with CBR, for MLP 78.85% with E-CBR and 73.59% with CBR). Moreover, when using CBR, KNN achieved better results than RF, SVM, C4.5 and MLP under MCAR and MAR regardless MD percentages (the mean balanced accuracy rates in average under MCAR are: for KNN: 81.20%, for C4.5: 77.63%, for RF 79.72%, for SVM: 79.35% and for MLP: 73.59%; under MAR: for KNN: 78.47%, for C4.5: 72.88%, for RF 77.80%, for SVM: 77.55% and for MLP: 72.27%). The results are statistically confirmed using the Wilcoxon and Holm-Bonferroni tests (under MCAR: KNN significantly outperformed SVM, C4.5 and MLP using either E-CBR or CBR: for C4.5, E-CBR: $p(\alpha) = 0.021/p(\alpha') = 0.025$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0166$; for SVM, E-CBR: $p(\alpha) = 0.008/\ p(\alpha') = 0.0125$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.025$; and for MLP, E-CBR: $p(\alpha) = 0.008/p(\alpha') = 0.0125$, CBR: $p(\alpha) = 0.008/p(\alpha') = 0.05$). However, under MCAR, KNN with CBR showed better results than SVM, MLP with E-CBR regardless the MD percentage (the mean balanced accuracy rates obtained in average for KNN with CBR and SVM, MLP with E-CBR are respectively: under MCAR: 81.20%, 80.73%, and 78.85%). While, under MAR KNN with CBR showed better results than MLP and C4.5 with E-CBR regardless the MD percentage (In average the mean balanced accuracy rates obtained under MAR for KNN with CBR and C4.5 and MLP with E-CBR are respectively: 78.47%,75.49 and 77.09%%), these results are confirmed by the Wilcoxon and Holm-Bonferroni statistic tests (under MCAR: C4.5: $p(\alpha) = 0.008/\ p(\alpha') = 0.0125$, SVM: $p(\alpha) = 0.008/p(\alpha') = 0.0166$, MLP: $p(\alpha) = 0.008/p(\alpha') = 0.0166$; under MAR: C4.5: $p(\alpha) = 0.008/p(\alpha') = 0.0025$, SVM: $p(\alpha) = 0.008/p(\alpha') = 0.0166$, MLP: $p(\alpha) = 0.008/p(\alpha') = 0.0166$).

The combination C4.5 with E-CBR showed better results than SVM and MLP under MCAR using E-CBR regardless the MD percentage (under MCAR the mean balanced accuracy rates achieved by C4.5, SVM and MLP using E-CBR in average are respectively: 81.23%, 80.74% and 78.52%). These results were confirmed by the Wilcoxon and Holm-Bonferroni statistical tests (Under MCAR: for SVM: $p(\alpha) = 0.086/p(\alpha') = 0.05$ and for MLP: $p(\alpha) = 0.0086/p(\alpha') = 0.0166$). Moreover, under MAR, MLP outperformed C4.5 regardless the MD percentage (In average, the mean balanced accuracy rates obtained by MLP and C4.5 under MAR are: with E-CBR: 77.09%and 75.49%. While under NMAR, MLP outperformed C4.5 and SVM regardless the MD percentage (In average, the mean balanced accuracy rates obtained by MLP, C4.5 and SVM under NMAR are respectively: with E-CBR: 80%, 78.011% and 79.35%; with CBR: 79.35%, 74.68% and 77.95%). These results were confirmed by the Wilcoxon and Holm-Bonferroni statistical tests (under MAR: for

C4.5 $p(\alpha) = 0.028/p(\alpha') = 0.025$, for SVM: $p(\alpha) = 0.066/p(\alpha') = 0.025$; under NMAR: for MLP $p(\alpha) = 0.004/p(\alpha') = 0.0166$, for SVM: $p(\alpha) = 0.004/p(\alpha') = 0.025$).

The combination SVM with either E-CBR or CBR achieved better results than MLP under MCAR regardless the MD percentage (Under MCAR, the mean balanced accuracy rates achieved by SVM and MLP in average are respectively: with E-CBR: 80.73% and 78.85%, with CBR: 79.35% and 73.59%). These results were confirmed statistically by the Wilcoxon and Holm-Bonferroni statistical tests (under MCAR: $p(\alpha) = 0.008/p(\alpha') = 0.025$ with CBR and $p(\alpha) = 0.008/p(\alpha') = 0.0125$ with E-CBR). While under MAR and NMAR, SVM outperformed C4.5 regardless the MD percentage (the mean balanced accuracy rates obtained by SVM and C4.in average under MAR are respectively: 78.52% and 75.49% with E-CBR; 77.55% and 72.88% with CBR). These results were confirmed by the Wilcoxon and Holm-Bonferroni statistical tests ($p(\alpha) = 0.008/p(\alpha') = 0.05$ with CBR and $p(\alpha) = 0.008/p(\alpha') = 0.0125$ with E-CBR).

According to Figs. 2-7, we summarize the findings of the RQs 1-2 as follow:

1. The RF classifier with E-CBR achieved the highest balanced accuracy rates regardless the MD mechanism, followed by KNN with E-CBR compared to SVM and C4.5 with E-CBR. This may be explained by the fact that using both ensembles in imputation and classification may achieve better results, since RF is an ensemble of DTs.
2. RF with CBR achieved better results than other classifiers with E-CBR. For instance, RF with CBR outperformed C4.5 with E-CBR under MCAR, MLP with E-CBR under MAR regardless the MD percentage.
3. In general, all classifiers tolerate higher percentages of MD when using both E-CBR and CBR, and it still yields acceptable balanced accuracy rates even if the percentage of MD is high.
4. It's noteworthy that in general MLP using CBR achieved the lowest balanced accuracy rate under MCAR and MAR regardless the MD percentage, while C4.5 using CBR achieved the lowest results under NMAR.

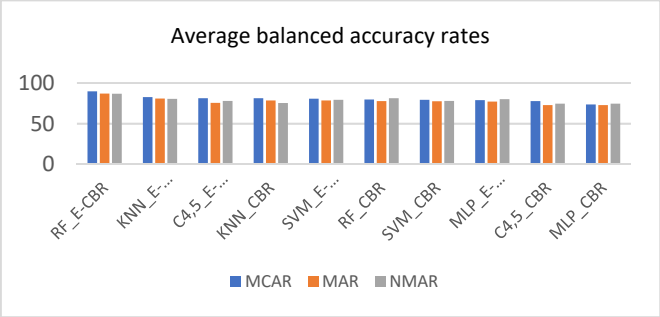

Fig. 7. Mean balanced accuracy rates of C4.5/KNN/RF/SVM/MLP using CBR and E-CBR, three MD mechanisms, and nine MD percentages.

4. CONCLUSION AND FUTURE WORK

This study aimed to evaluate the impact of CBR based ensemble imputation of MD on the performances of five classifiers: C4.5, KNN, RF, SVM and MLP over two breast

cancer datasets. The performance of each classifier was evaluated in terms of the balanced accuracy criterion using three MD mechanisms (MCAR, MAR and NMAR) with nine percentages (from 10% to 90%). Thereafter, we compared the ensemble based CBR imputation with a single CBR using GS. The findings proved that the classification performance achieved when using the E-CBR technique outperformed the performance with CBR for the five classifiers. Therefore, we can conclude that the use of ensemble imputation instead of GS single imputation improved significantly the accuracy of BC classification regardless the MD mechanism and percentage. Moreover, RF using E-CBR yield to better accuracy rates compared to the other classifiers regardless the MD mechanism and percentage.

Ongoing research intends to investigate other imputation ensembles: homogenous and heterogeneous based on other single imputers such as SVM and DTs.

AKNOWLEDGEMENT

This work was conducted under the research project "Machine Learning based Breast Cancer Diagnosis and Treatment", 2020-2022. The authors would like to thank the Moroccan Ministry of Higher Education, ADD, CNRST, and UM6P for their support.

REFERENCES

1. R. J. Oskouei, N. M. Kor, and S. A. Maleki, "Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges," *American Journal of Cancer Research*, Vol. 7, 2017, pp. 610-627.
2. A. Idri, I. Abnane, and A. Abran, "Support vector regression-based imputation in analogy-based software development effort estimation," *Journal of Software: Evolution and Process*, Vol. 30, 2018, pp. 1-23.
3. B. Garg, "Optimizing number of inputs to classify breast cancer using artificial neural network," *Journal of Computer Science and Systems Biology*, Vol. 2, 2009, pp. 247-254.
4. I. Kadi, A. Idri, and J. L. Fernandez-Aleman, "Knowledge discovery in cardiology: A systematic literature review," *International Journal of Medical Informatics*, Vol. 97, 2017, pp. 12-32.
5. A. Idri, H. Benhar, J. L. Fernández-Alemán, and I. Kadi, "A systematic map of medial data preprocessing in knowledge discovery," *Computer Methods and Programs in Biomedicine*, Vol. 162, 2018, pp. 69-85.
6. N. Esfandiari, M. R. Babavalian, A. M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems With Applications*, Vol. 41, 2014, pp. 4434-4463.
7. M. Ezzat and A. Idri, "Reviewing data analytics techniques in breast cancer treatment," in *Trends and Innovations in Information Systems and Technologies*, 2020, pp. 65-75.
8. A. Idri, I. Chlioui, and B. El Ouassif, "A systematic map of data analytics in breast cancer," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2018, pp. 1-10.
9. B. P. Vrigazova, "Detection of malignant and benign breast cancer using the ANOVA-BOOTSTRAP-SVM," *Journal of Information Science*, Vol. 5, 2020, pp. 62-75.

10. M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare*, Vol. 8, 2020, p. 111.
11. E. Acuña and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Proceedings of the Meeting of the International Federation of Classification Societies*, 2004, pp. 639-647.
12. I. Chlioui, A. Idri, I. Abnane, C.-G. J. Manuel, Alemán, and J. L. Fernández, "Breast cancer classification with missing data imputation," *New Knowledge in Information Systems and Technologies*, 2019, pp. 13-23.
13. I. Chlioui, I. Abnane, and A. Idri, "Comparing statistical and machine learning imputation techniques in breast cancer classification," in *Proceedings of International Conference on Computational Science and Its Applications*, 2020, pp. 61-76.
14. I. Chlioui, A. Idri, and I. Abnane, "Data preprocessing in knowledge discovery in breast cancer: Systematic mapping study," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 2020, pp. 1-15.
15. A. Idri, I. Abnane, and A. Abran, "Missing data techniques in analogy-based software development effort estimation," *Journal of Systems and Software*, Vol. 117, 2016, pp. 595-611.
16. S. Jhajharia, H. K. Varshney, S. Verma, and R. Kumar, "A neural network based breast cancer prognosis model with PCA processed features," in *Proceedings of International Conference on Advances in Computing*, *Communications and Informatics*, 2016, pp. 1896-1901.
17. X. Ma, Y. Zhang, and Y. Wang, "Performance evaluation of kernel functions based on grid search for support vector regression," in *Proceedings of the 7th IEEE International Conference on Cybernetics and Intelligent Systems*, 2015, pp. 283-288.
18. S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, Vol. 1, 1997, pp. 317-328.
19. V. García, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," in *Proceedings of Iberian Conference on Pattern Recognition and Image Analysis*, 2009, pp. 441-448.
20. K. Kafadar and D. J. Sheskin, "Handbook of parametric and nonparametric statistical procedures," *The American Statistician*, Vol. 51, 2006, p. 374.
21. H. Abdi, "1 Overview 2 Preliminary : The different meanings of alpha," *Encyclopedia of Research Design*, Vol. 51, 2006, pp. 1-8.
22. The University of Waikato, "Weka the university of Waikato" https://www.cs.waikato.ac.nz/ml/weka/.
23. L. Peng and L. Lei, "A review of missing data treatment methods," *Information Intelligence*, *Systems*, *Technology and Management*, Vol. 1, 2005, pp. 1-8.
24. A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Optimal K parameter for KNN classifier with square root," *International Journal of Computer Science and Information Security*, Vol. 12, 2014, pp. 33-39.

Imane Chlioui is a Ph.D. student at the Computer Science and Systems Analysis School (ENSIAS, University Mohammed V, Rabat, Morocco), a member of the Software Project Management Research Team. Her doctoral research investigates the influence of missing data techniques on breast cancer classification. She received her engineering degree in 2015 in Computer Science from the Computer Science and Systems Analysis School.

Ali Idri is a Full Professor at the Computer Science and Systems Analysis School (ENSIAS, University Mohammed V, Rabat, Morocco). He received his Master and Doctorate of 3rd Cycle in Computer Science from the University of Mohamed V in 1994 and 1997 respectively. He received his Ph.D. in Cognitive and Computer Sciences from the University of Quebec at Montreal in 2003. He is the Head of the Software Project Management Research Team since 2010. He published more than 200 papers in well-recognized journals and conferences. His research interests include medical informatics, machine learning and software engineering.

Ibtissam Abnane is an Assistant Professor at the Computer Science and Systems Analysis School (ENSIAS, University Mohammed V, Rabat, Morocco). She received her engineering degree in Computer Science from National School of Applied Sciences of Safi (ENSAS) in 2013. She received her Ph.D. in Computer Sciences from the Computer Science and Systems Analysis School (ENSIAS, University Mohammed V, Rabat, Morocco) in 2018. She is a member of the Software Project Management Research Team. She is working in the fields of software engineering, machine learning and medical informatics.

Mahmoud Ezzat is an Engineer from the highest school in Computer Sciences in Rabat (ENSAS). Currently, he is working as a Data Scientist within DFR entity and he also pursuits a Ph.D. degree in MSDA Department, Mohammed VI Polytechnic University. His research interests include machine learning and breast cancer treatment.