Breast Cancer Classification with Missing Data Imputation

Imane Chlioui[1], Ali Idri[1(✉)], Ibtissam Abnane[1], Juan Manuel Carillo de Gea[2], and Jose Luis Fernández-Alemán[2]

[1] Software Project Management Research Team, ENSIAS, University Mohammed V of Rabat, Rabat, Morocco
ali.idri@um5.ac.ma

[2] Department of Informatics and Systems, Faculty of Computer Science, University of Murcia, Murcia, Spain

Abstract. Missing Data (MD) is a common drawback when applying Data Mining on breast cancer datasets since it affects the ability of the Data mining classifier. This study evaluates the influence of MD on three classifiers: Decision tree C4.5, Support vector machine (SVM), and Multi-Layer Perceptron (MLP). For this purpose, 162 experiments were conducted using KNN imputation with three missingness mechanisms (MCAR, MAR and NMAR), and nine percentages (form 10% to 90%) applied on two Wisconsin breast cancer datasets. The MD percentage affects negatively the classifier performance. MLP achieved the lowest accuracy rates regardless the MD mechanism/percentage.

Keywords: KNN imputation · Data mining · Breast cancer

1 Introduction

In the past twenty years, the number of patients with breast cancer (BC) continues to rise, and it became the second leading cause of death among women [1]. It is a malignant tumor that has developed from cells in the breast [2]. The key to an effective treatment is early diagnosis: the earlier the disease is diagnosed the less it progresses. Nowadays, the amount of data is increasing constantly in all fields such us: education, agriculture and medicine [3]. Which the necessity of the use of data mining (DM) techniques to analyze the huge amount of available data and extract knowledge [1]. According to Idri et al. [4] the use of DM techniques has increased lately, and become a powerful tool to help radiologists and practitioners to deal with BC challenges.

To successful the use of DM techniques, the preprocessing step is recommended to avoid biased and deceptive results. Cleaning, transformation, reduction, and integration are subfields of preprocessing. Handling missing data (MD) as a part of the cleaning process is a major problem facing the use of DM tools [5]. Thus, several MD techniques have been proposed and experimented; they can be grouped in three categories [6, 7]: (1) toleration technique which consists on ignoring the MD, (2) deletion technique which consists on deleting the MD, and (3) imputation techniques which consist on filling in the MD with appropriate values.

Á. Rocha et al. (Eds.): WorldCIST'19 2019, AISC 932, pp. 13–23, 2019.
https://doi.org/10.1007/978-3-030-16187-3_2

Therefore, this paper analyses and discusses the impact of the use of KNN-imputation on the accuracy of three classifiers: decision tree C4.5, support vector machine (SVM) and multi-layer perceptron (MLP) over two datasets: Wisconsin breast cancer original and Wisconsin breast cancer prognosis. Moreover, the empirical evaluations used three MD missingness mechanisms (MCAR, MAR, NMAR), nine MD percentages (from 10% to 90%), and were performed using the experimental process proposed by Idri et al. [7]. To the best of our knowledge, no existing study that analyzes the impact of KNN imputation using different MD mechanisms (MCAR, MAR, and NMAR) with nine percentages (from 10% to 90%) on the performance of classification techniques in breast cancer, which motivates this study.

This paper is structured as follows. Section 2 introduces the different MD mechanisms and MD imputation techniques. Related work dealing with missing values in breast cancer is presented in Sect. 3. Section 4 introduces the datasets as well as the classification techniques used. The experimental design followed in this study is detailed in Sect. 5, while the Sect. 6 presents the results and discuss the findings. Threats to validity are presented in Sect. 7. Section 8 concludes the paper and proposes further research lines.

2 Missing Data Concepts

The MD mechanisms and technique used in this study are presented in the section follow.

2.1 Missing Data Types

The missingness mechanism indicates the reason of data missingness and could help to choose the suitable MD technique. Three MD mechanisms where defined by Rubin [8]:

Missing Completely at Random (MCAR): MD are independent of other variables and there is no specific reason of missingness [9].

Missing at Random (MAR): MD is not related to the missing values themselves, but is related to other observed variables. This mechanism can bias the results and may cause unbalanced data [10].

Not Missing at Random (NMAR): MD are dependent to the MD themselves. This can happen when the variable is not observed or it takes a value out of its representation range. This MD type may give highly biased estimation results [10].

2.2 Missing Data Imputation Techniques

In contrast to deletion technique that permits to discard instances that contain missing values, imputation techniques replace missing items with plausible values [11]. KNN imputation has been widely used by several researchers, and it is performed by considering the K closest instances to the incomplete instance according to a given distance

metric. Several distance measures were proposed in literature such as: Euclidian distance, Manhattan distance and Hamming distance [12]. This study employs the Euclidean distance which assesses the distance between instances x_i and y_i by the Eq. (1).

$$D_E(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Where n is the number of attributes describing the instances x_i and y_i.

3 Related Work

This section presents a summary of two papers selected from the systematic mapping study of [4].

García-Laencina et al. [13] applied several MD techniques on a breast cancer dataset with a high percentage of MD in order to predict breast cancer survivability. The dataset used in this study is collected from the Institute Portuguese of Oncology of Porto (IPO). They found that KNN imputation was the best in terms of accuracy, while the Mode imputation was the worst.

Jerez et al. [14] Compared statistical/machine learning imputation techniques with deletion to predict the breast cancer prognosis. The study proved that all imputation techniques except Hot-deck could improve the accuracy of an artificial neural network (ANN) based prediction. Moreover, ML imputation techniques outperformed statistical ones and led to statistically significant improvements in prediction accuracy. The best predictions were obtained using the KNN imputation with an improvement of 2.71% over deletion.

4 Datasets and Classifiers Description

This section presents a brief description of the datasets used and the classification techniques investigated.

4.1 Datasets Description

The datasets used in this study were collected at the University of Wisconsin–Madison Hospital [25]. The first one is the Wisconsin breast cancer original dataset, most commonly employed by researchers investigating machine learning techniques for breast cancer. It contains 699 instances described by 10 numerical attributes. The second one is the Wisconsin breast cancer prognosis dataset; it contains 198 records descried by 35 numerical attributes.

All cases containing missing data were discarded, which reduces the size of each dataset: 683 in Wisconsin original and 194 in Wisconsin prognosis. Moreover, the attributes of Wisconsin breast cancer prognosis dataset were normalized within the interval [0–1] 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, along with the number of instances and attributes. All the attributes used in this study are numerical.

Table 1. Datasets description

Database	Instances	Attributes	Attributes type	Source
Wisconsin breast cancer original	194	35	Numeric	[15]
Wisconsin breast cancer prognosis	683	10	Numeric	[16]

4.2 Classification Techniques

Hereafter, we describe the three classifiers used in this study.

C4.5. A supervised learning classification algorithm used to construct decision trees from the data developed by Quinlan [17].

SVM. A group of supervised learning methods developed by Vapnik in the 90's [18]. It is used to model data not linearly separable [19, 20].

MLP. A type of artificial neural network (ANN) that can represent complex input-output relationships [21]. 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 output.

5 Empirical Design

Figure 1 presents the empirical process we used in this study. We followed the same empirical process used by [7] to handle missing data in software development effort estimation. This process consists of four main phases: data removal, complete dataset generation using imputation techniques, application of classification techniques, and accuracy evaluation. Each step of this process is detailed in the following subsections.

5.1 Data Removal

The datasets should be complete to work with. For this reason, the two datasets were cleaned by deleting all the cases with MD. Thereafter, the MD were generated artificially using the three missing data mechanisms:

- The MCAR mechanism relies on the randomization; the MD was induced completely at random for each variable.
- The MAR mechanism was simulated relying on a single attribute of each dataset: cell_shape_uniformity for Wisconsin original dataset and lymph_node_status for Wisconsin prognosis dataset. First, the instances were sorted in an ascending order of the selected attribute. Thereafter, the data were split into three equal subsets. The MD were distributed as follows: (1) 60% * p assigned randomly to the first subset, (2) 40% * p assigned to the second subset, and (3) 0% to the third subset; p

is the percentage of MD. The MD were induced to all attributes with bias related to the two selected attributes.

- The NMAR mechanism is similar to the MAR, except that the Wisconsin original dataset was sorted according to cell_size_uniformity. The only difference between NMAR and MAR mechanisms is that for NMAR, MD were not induced to all attributes but only to the attribute to which the datasets were sorted (i.e. cell_-size_uniformity and lymph_node_status).

For each missingness mechanism, 9 percentages (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%) were simulated, which gave us a total of 54 incomplete datasets (54 = 3 MD mechanisms * 9 percentages * 2 datasets).

5.2 Complete Dataset Generation

In this step, KNN imputation was applied on the datasets resulted from the data removal step. The Euclidean distance was used to evaluate the similarity between instances since all the attributes are numerical. Thereafter, the median of the closest instances was used to fulfill the missing values. At the end of this step, 54 complete datasets were willing to be used for the classification task (54 = 54 incomplete datasets * 1 MD technique).

5.3 Generating Classifiers

For the classification task, three classifiers were applied on the complete datasets: C4.5, SVM, and MLP. To fulfill this task, the datasets were divided into training and testing sets using the 10-fold cross validation. After applying the three classifiers to the 54 data sets of the complete dataset generation step, we obtained 162 classification experiments (162 = 54*3). For each classifier, the grid search method was used to vary the classifiers parameters.

5.4 Performance Evaluation

To evaluate the performance of the three classifiers, the accuracy measure was used; it represents the probability of correctly predicting the class of an instance [22]. The accuracy rate was evaluated using the Eq. (2). TP or true positives is the number of positives cases correctly classified, TN or true negatives is the number of negatives cases correctly classified, FP or false positives, the number of positives cases classified as negatives, and FN or false negatives, the number of negatives cases classified as positives [23].

$$\text{Accuracy rate} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (2)$$

6 Results and Discussion

This section presents and discusses the empirical results when applying the three classifiers using KNN imputation, three MD mechanisms, and nine percentages over the two Wisconsin datasets. Therefore, we analyze the impact of each MD mechanism on the classifiers accuracy under different MD percentages. All the empirical evaluations were implemented using the WEKA (3.8.0) tool [26].

In order to determine the best performance of each classifier, the grid search method was used to set up the ranges of parameters and test each combination to find the best one [27]. The optimal configuration (i.e., that which attained the best performance results) of each classifier was then used in the subsequent experiment. Table 2 presents the parameter ranges of each classifier and the optimal configuration of each classifier as well.

6.1 Evaluation of Classifiers Using MD Mechanisms

This section presents and discusses the influence of the three missingness mechanisms on the mean accuracy rates of the three classifiers using KNN imputation and nine percentages.

- For the SVM classifier using KNN imputation, Fig. 2a shows that the mean accuracy rates obtained under MCAR were better than under MAR and NMAR regardless the MD percentages (at 10% of MD the mean accuracy rates were: under MCAR 90.10%, under MAR 88.26% and under NMAR 88.30%). Moreover, under NMAR the mean accuracy rates were higher than those obtained under MAR regardless the MD percentages (at 90% of MD the mean accuracy rates were: under NMAR 88.16% and under MAR 87.31%.

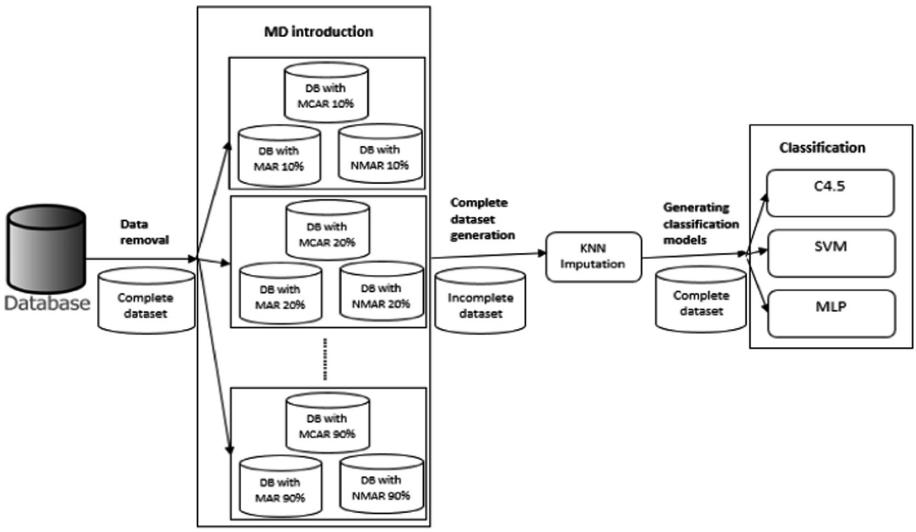

Fig. 1. Experimental design

Table 2. Parameters ranges and optimal configuration of each classifier

Algorithm	Parameters ranges	Optimal configuration
C4.5	C = {0.1 –> 5, increment = 0.1}; M = {10 –> 100, increment = 10};	C = 0.25 M = 2
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
KNN imputation	K = {2 –> 7, increment = 1}	K = 5

- For the MLP classifier using KNN imputation, according to Fig. 2b the mean accuracy rates achieved under MAR were better than under MCAR and NMAR regardless the MD percentages (at 10% of MD the mean accuracy rates were: under MAR 87.82%, under MCAR 86.53% and under NMAR 84.77%). Furthermore, under MCAR the mean accuracy rates were higher than those obtained under NMAR (at 90% of MD the mean accuracy rates were: under MCAR 86.02%, and under NMAR 81.19%).
- For the C4.5 classifier using KNN imputation, according to Fig. 2c the mean accuracy rates realized under MAR were better than under MCAR and NMAR regardless the MD percentages (at 10% of MD the mean accuracy rates were: under MAR 89.80%, under NMAR 88.11% and under MCAR 88.08%). Furthermore, under NMAR the mean accuracy rates were higher than those obtained under MCAR (at 90% of MD the mean accuracy rates were: under NMAR 88.04%, and under MCAR 85.61%).

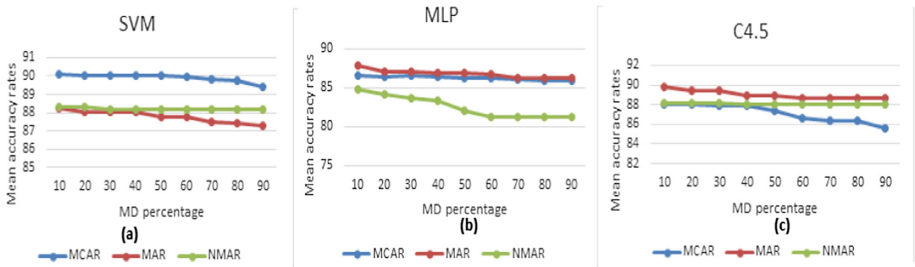

Fig. 2. Mean accuracy rates of SVM, MLP and C4.5 using KNN imputation with three MD mechanisms and nine MD percentages.

6.2 Comparison of Classifiers Using MD Mechanisms

This section compares the mean accuracy rates of the three classifiers C4.5, SVM and MLP, using KNN imputation, three MD mechanisms, and nine MD percentage. Figure 3a–c show the mean accuracy rates of each classifier using KNN imputation, with three MD mechanisms and nine MD percentages.

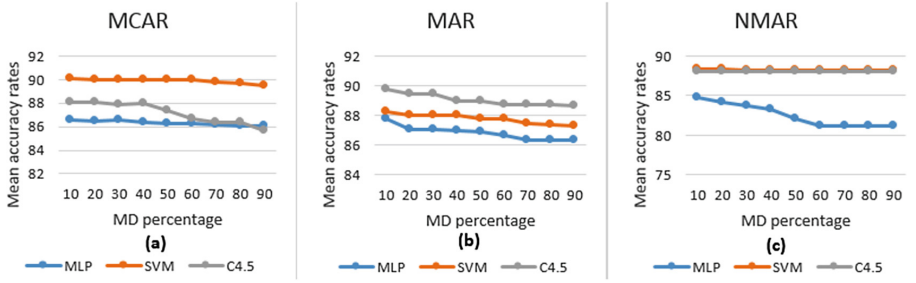

Fig. 3. Comparison of three classifiers using KNN imputation with three MD mechanisms and nine MD percentages

As can be seen in Fig. 3a–c, SVM achieved the highest accuracy rates followed by C4.5 under MCAR and NMAR regardless the MD percentage (for example at 10% of MD the accuracy rates obtained under MCAR are: 90.10% for SVM, 88.08% for C4.5 and 86.53% for MLP). While under MAR C4.5 achieved better results than SVM and MLP regardless the MD percentage (for example at 10% of MD the accuracy rates obtained under MAR are: 89.80% for C4.5, 88.26% for SVM, and 87.88% for MLP).

It's noteworthy that MLP achieved the lowest accuracy rates regardless the MD mechanisms and percentage.

According to Figs. 2a–c and 3a–c, we summarize the findings:

1. The MD percentage has an important influence on the accuracy rates of the classifiers, as long as the MD percentage increases the accuracy rate decreases; which can be explained by the fact that 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 [24].
2. SVM showed the better accuracy rates comparing to C4.5 and MLP, and this can be explained by the fact that SVM proved helpful in breast cancer diagnosis [22]. The imputation can perform better for SVM while C4.5 is more missing data resistant, and can handle missing data [23].
3. Although ANNs are often considered an advanced machine learning and very sophisticated, MLP showed the lowest results in our study regardless the MD mechanism/percentage.

7 Threats to Validity

Internal validity: Internal threats of this paper are in general corresponding to the evaluation of the classifiers. The first one is related to the evaluation metric used; the results of this study were based on the accuracy rate metric because it is widely used to evaluate the classification performance. The second threat is related to the evaluation method, 10 cross-validation model was adopted because it is considered as a standard for performance estimation and technique selection [24].

External validity: this study used only numerical attributes. Further investigations on other datasets are required to discuss categorical attributes. In this empirical evaluation, only three classifiers were applied to investigate the impacts of MD mechanisms. Thus, more classifiers should be evaluated in order to assess the impact of MD. Moreover, there is several imputation techniques used to handle MD; in this work only KNN imputation was used due to its popularity, yet other imputation techniques can achieve better results.

8 Conclusion and Future Work

In this study, the impacts of MD on the three classifiers: C4.5, SVM and MLP were evaluated over two datasets from the UCI repository, the Wisconsin breast cancer original and prognosis datasets. The performance of each classifier was evaluated using KNN imputation with three MD mechanisms (MCAR, MAR and NMAR) and nine percentages (from 10% to 90). According to the obtained results, the missingness mechanism influences the accuracy rates of the classifiers, and varies for different classifiers. On the other hand, the MD percentage influences negatively the classifiers accuracy, while the MD percentage increases the accuracy rate decreases regardless the MD mechanism and technique. SVM yelled to better results under MCAR and NMAR regardless the MD percentage, while MLP achieved the lowest accuracy rates.

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 compare the impact of KNN imputation technique and deletion on more classifiers such us Random Forest (RF) and Case-based reasoning (CBR). Since the present study only deals with numerical attributes, it would be of great interest to deal with missing categorical data.

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. (2017)
2. Akay, M.F.: Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. **36**, 3240–3247 (2009). https://doi.org/10.1016/j.eswa.2008.01.009

3. Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., Tabar, V.K.: Knowledge discovery in medicine: current issue and future trend. Expert Syst. Appl. (2014). https://doi.org/10.1016/j.eswa.2014.01.011
4. Idri, A., Chlioui, I., Ouassif, B.E.: A systematic map of data analytics in breast cancer. In: ACSW 2018 Proceedings pf Australasian Computer Science Week Multiconference, Brisband, pp. 26:1–26:10 (2018). https://doi.org/10.1145/3167918.3167930
5. Cismondi, F., Fialho, A.S., Vieira, S.M., Reti, S.R., Sousa, J.M.C., Finkelstein, S.N.: Missing data in medical databases: impute, delete or classify? Artif. Intell. Med. (2013). https://doi.org/10.1016/j.artmed.2013.01.003
6. 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). https://doi.org/10.1016/j.cmpb.2018.05.007
7. 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
8. Rubin, D.B.: Inference and missing data (with discussion). Biometrika **63**, 581–592 (1976)
9. Garciarena, U., Santana, R.: An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst. Appl. **89**, 52–65 (2017). https://doi.org/10.1016/j.eswa.2017.07.026
10. Curley, C., Krause, R.M., Feiock, R., Hawkins, C.V.: Dealing with missing data : a comparative exploration of approaches using the integrated city sustainability database (2017). https://doi.org/10.1177/1078087417726394
11. 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
12. Yenduri, S.: An empirical study of imputation techniques for software data sets (2005)
13. García-Laencina, P.J., Abreu, P.H., Abreu, M.H., Afonoso, N.: Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput. Biol. Med. **59**, 125–133 (2015). https://doi.org/10.1016/j.compbiomed.2015.02.006
14. Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L.: 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
15. Index of /ml/machine-learning-databases/breast-cancer-Wisconsin (2017). Archive.ics.uci.edu. https://ww.archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). Accessed 20 Jul 2003
16. Index of /ml/machine-learning-databases/breast-cancer-wisconsin (2017). Archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(Prognostic). Accessed 20 Jul 2003
17. 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. (2008). https://doi.org/10.1016/j.jss.2008.05.008
18. Hall, M., Witten, I., Frank, E.: Data Mining, 4th Edn., Elsevier (2011)
19. Alpaydın, E.: Introduction to Machine Learning, 2nd Edn., The MIT Press, London (2014). https://doi.org/10.1007/978-1-62703-748-8-7
20. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel based learning methods. Cambridge University Press, Cambridge (2000). citeulike-article-id:114719

21. Ghosh, S., Mondal, S., Ghosh, B.: A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In: 2014 First International Conference on Automation, Control, Energy and System, pp. 1–4 (2014). https://doi.org/10.1109/aces.2014.6808002
22. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics (2000). https://doi.org/10.1093/bioinformatics/16.5.412
23. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. (2006). https://doi.org/10.1016/j.patrec.2005.10.010
24. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Min. Knowl. Discov. (1997). https://doi.org/10.1023/a:1009752403260
25. 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, pp. 1896–1901 (2016). https://doi.org/10.1109/ICACCI.2016.7732327
26. The university of Waikato, Weka the university of Waikato, (n.d.). https://www.cs.waikato.ac.nz/ml/weka/
27. Ma, X., Zhang, Y., Wang, Y.: Performance evaluation of kernel functions based on grid search for support vector regression. In: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems and IEEE Conference on Robotics, Automation and Mechatronics, pp. 283–288 (2015). https://doi.org/10.1109/ICCIS.2015.7274635