Towards an Oversampling Method to Improve Hepatocellular Carcinoma Early Prediction

Mahbouba Hattab, Ahmed Maalel, and Henda Hajjami Ben Ghezala

1 Introduction

Knowledge discovery in medical databases has become an attractive and crucial complement for clinical research. Survival and disease prediction are of a highly important task addressed by the medical research communities due to its direct effect on doctor's decisions [1]. Using KDD (Knowledge Discovery in Databases), efficient and important knowledge can be extracted from these data sets. The principal steps in the KDD process are as follows: (1) Data selection, (2) Data management and pretreatment, (3) Transformation, (4) Data mining, and (5) Assessment and interpretation. Ideally, from a computer science perspective, data mining is one of the ultimate steps in the KDD process. Indeed, Data Mining is a discipline resulting from combining statistics and computer science such as Machine Learning algorithms. Data mining aims to extract new and useful knowledge from a large amount of data (i.e.: applied to have an effective and preferment predictive model [2]). However, each of the aforementioned domains has its specifics, therefore it is

M. Hattab (✉)
Higher Institute of Applied Sciences and Technology of Sousse, University of Sousse, Sousse, Tunisia

A. Maalel
Higher Institute of Applied Sciences and Technology of Sousse, University of Sousse, Sousse, Tunisia

National School of Computer Sciences, RIADI Laboratory, University of Manouba, Manouba, Tunisia
e-mail: ahmed.maalel@ensi.rnu.tn

H. H. B. Ghezala
National School of Computer Sciences, RIADI Laboratory, University of Manouba, Manouba, Tunisia

L. Chaari (ed.), *Digital Health in Focus of Predictive, Preventive and Personalised Medicine*, Advances in Predictive, Preventive and Personalised Medicine 12,
https://doi.org/10.1007/978-3-030-49815-3_16

important to wisely choose the best data optimization and pre-treatment algorithm to achieve state-of-the-art classification accuracy.

Due to the abundance of data in the biomedical domain, this latter has a high potential in improving the well being of humankind. Nevertheless, it is very complex to process and analyze such data by traditional methods [1]. As a result, the interest of data mining and machine learning is increasing considerably with a wide range of medical applications. They become useful instruments in bioinformatics thanks to their capability to convert this vast resource into information and knowledge that helps achieve better decision making in several disease areas including cardiovascular disorders, Parkinson's, Alzheimer's, etc. [3].

Cancer is a generic term for a large variety of diseases that can affect any person and any part of the body and it represents one of the leading causes of morbidity and mortality worldwide after heart diseases [4, 15]. Globally, according to the World Health Organisation (WHO) [4], one in six deaths are due to cancer. Relatable to cancer, Hepatocellular carcinoma (HCC) also referred to as malignant hepatoma [3] is a malignant tumor, and represent the sixth most common type of cancer and the third leading cause of cancer-related deaths globally according to [5].

2 Related Works and Motivation

Over the past years, several research works have been conducted on HCC and liver-related diseases. For instance, **Santos et al. (2015)** [5] studied HCC data set using a new cluster-based oversampling algorithm. The proposed methodology is based on the data pre-treatment process considering appropriate distance metrics for both heterogeneous and missing data by applying the Heterogeneous Euclidean-Overlap Metric (HEOM) distance. Then Kmeans clustering algorithm is applied for the first sampling step within the HCC database and SMOTE oversampling algorithm to build a representative balanced data set and use it for Leave-One-Out cross Validation (LOO-CV) assessment with different machine learning algorithms such as logistic regression (LR) and neural networks (NN) classifiers. The results indicated that the proposed approach can achieve efficient results.

In another work, **Sawhney et al.** (2018) [6] explored the performance of the firefly algorithm by adding a penalty function to the existing fitness function. Afterward, they modify the existing wrapper feature to reduce the feature set to an optimal subset. Furthermore, the influence of the method is proved on the classification accuracy as well as feature reduction using a Random Forest classifier for the Hepatocellular Carcinoma dataset in comparison to other contemporary methods such as Deep Learning methods and Information Gain. However, the above-mentioned works, are of a highly computationally expensive and require a large amount of data in order to generate a satisfying model. Moreover, real-world data tends to be incomplete, noisy, and inconsistent and the important task is to fill in missing values, smooth out noise and correct inconsistencies. To address these issues, previous works have relied on applying feature selection [7] to eliminate the

redundant and inconsistent data and thus improve the capacity of the classifier. Other studies used multiple scaling methods such as normalization and standardization to improve the classification model [5, 6]. Despite the good *theoretical* approach of these methods, the delivered results can be tremendously inconsistent as each of the feature selection or the normalization/standardization methods can deliver a different result and thus selecting the best approach is a time consuming and exhausting process. Another important issue that could be found which is:

- outliers: "*Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism"* Hawkins(1980)

In Data Science, an Outlier is an observation point that is distant from other observations on data that diverges from an overall pattern on a sample. They may indicate variability in measurement, experimental errors or a novelty. The quality and the prediction speed of classifier depend on the input data set for the training thus, the more outliers existing in the training set is, the less accurate the prediction is [7, 8]. In fact the most common causes of outliers on a data set are: (1) Incorrect data entry, (2) Data processing errors: application of inappropriate missing values methods in a dataset, (3) Outliers case did not come from the intended sample and (4) Not an outlier, just a novelty in data.

Instance selection is a recommended technique that was developed to overcome the limitations related to noise and outliers. The aim behind using instance selection is to improve the prediction accuracy of the classifier. To achieve that goal these algorithms are designed to remove outliers and noisy instances. The ensemble learning has become one of the most promising machine learning approaches during the last decade. It takes advantage of combining several models, which when grouped together, can outperform each method with only a linear increase in computational complexity [8]. Several states of the art were included in these groups of algorithms. We include, in the first one, an ensemble of C different models trained on the same training set of data T vote for the output of the instance being classified t. In another state of the art, an extension of the voting approach is presented; where c independent models of different types are trained on the same data set T, then all models outputs are combined for an extra model for the final prediction. Also, a third approach consists of c models of the same type, where each of them is trained on the data set T* that is obtained from T by sampling. The final prediction is obtained by voting. Finally, a similar approach to the previous, but the probability of selecting an instance from T depends on a classification error of the previous models so that an instance that was incorrectly classified by the current models is more likely to be selected.this method called Boosting [8].

For small data sets, noise filtering can be used as a step in the pre-processing of training data; on the other hand, the sampling process which selects a subset of the training set may include or followed by noise and outliers filtering.

In machine learning, Data clustering is frequently used in many fields, such as sampling of data (Stratified sampling) [9]. The mechanism of dividing a set of data to a particular set of groups is termed as clustering, and one such prominent methodology is the k-means clustering. Due to the nature of our available data, its

efficiency and its usefulness in several fields of pattern recognition particularly for clustering cancer data, K-means is a well-known unsupervised learning algorithm for data partition with a low computational cost. K-means iteratively reduces the Sum of Squared Error (SSE) from every object to their cluster centroids for every cluster Ck. (SSE) indicates the compactness of the cluster, the lower is, the better is.

Generally, to achieve success ensemble modeling, it recommended ensuring the diversity of the obtained results of each model, which are then combined into a final predictive one. The diversity of the results can be attained in several ways:

– Guarantee diversity of the models: Several different algorithms are used or the same algorithms but with different hyper-parameters or setup.
– Guarantee diversity of the data: we keep the same model, and each subset of the data used for training generates diversity.

In this paper, we address the issue of existing of outliers and noise in small datasets, so that data sampling, the instance selection and data oversampling are combined to ensemble methods and the diversity is obtained by manipulating different algorithms such as K-mean clustering and SMOTE. Our goal is that the ensemble of instance selection algorithm and the Oversampling technique allow us to manipulate the trade-off between the problem of small data sets and prediction accuracy. Moreover, improve both which grouped can outperform each method without an increase of computational complexity. This paper is organized as follows: the next section describes the basic algorithms used in the experiments. The following section describes the testing environment and presents numerical experiments. The last section summarizes the results and presents their interpretation.

3 Model Description

In the presented section, the different stages that compose the followed methodology to construct our approach, as well as the approach itself, are described.

First of all Fig. 1 represent the methodology applied in this work: Data imputation, Data partitioning, Clustering and finally classification. The main aspects of each stage are briefly described.

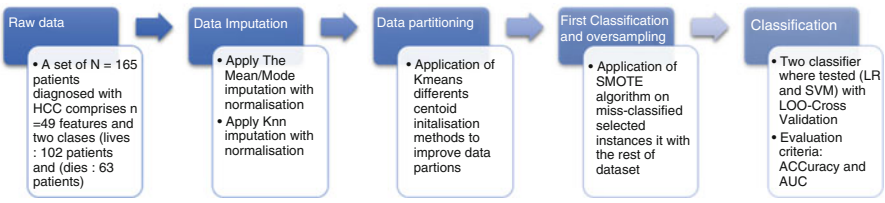

Fig. 1 Followed methodology

We used a HCC dataset composed of N = 165 patients diagnosed comprises n = 49 features provided by UCI machine learning repository.[1] The dataset's class distribution presents 63 cases labeled as 0 (dead) and the remaining 102 cases as 1 (alive) [5]. In the Data Imputation phase, two well-known imputation techniques (Mean/mode and K nearest neighbor) are used to handle missing values with Normalization. According to [10] Cross-validation is a common way of avoiding over-training. Nonetheless, the fundamental problem with this method is the proper partition of data. Simple random sampling is used for most applications. However, a variety of sophisticated methods of statistical sampling suitable for different types of datasets are available. The stratified sampling is one of these methods. The fundamental idea is to investigate the internal structure and distribution of the T dataset and to split T into relatively homogeneous sample groups. The samples were selected from each cluster separately [9]. Various clustering algorithms can be used to divide dataset T into clusters including K-means. For instance, The data sampling phase, in this work, aims to partition the initial data set into a set of tow different partitions by applying classification via clustering method with K-means algorithm for different seed values and three different centroid initialization methods (Random; Kmeans++; Canopy and Farthest first) to ensure diversity of the data. After the data partitioning phase, different resulting partitions are presented and the selection process was based mainly on the SSE that indicates how compact a cluster is: The lower the value, the better. then a manual review is applied to the chosen partitions: correctly labeled instances are kept to learn the classifier, then miss-labeled are used as a supplied test set for the next phase. Once final samples are produced, Our proposed method for selecting the outliers, which is based on the classification error of the previously built model, is performed and the resulting set is oversampled With the SMOTE algorithm. Thus, the oversampled partition is added to the initial dataset and duplicates instances are removed. The final phase consists of fit the classifier with the new augmented dataset. tow classifiers were engaged in this step, which are Logistic regression [11] and Support Vector Machine [12].

Our proposed approach is presented in Fig. 2. It starts by loading the data, then the data imputation is applied on missing values followed by Normalization for features scaling. After the pre-treatment phase, the stratified sampling with k-means is processed with different centroid initialization methods to find better subsets. For the following, and for rigorous scalability, reproducibility, and generalizability, the rest of the process was wrapped by the leave one out cross-validation for underlying subsets selection and the prediction process. It's important to note that the cross-validation strategy guarantees the best performance of the model [13] Step (C) is repeated several times in order to refine the resulting subset. Since the data set is relatively small, each instance is important for the prediction task.

[1] UCI Machine learning Repository, URL: https://archive.ics.uci.edu.

4 Experimental Results and Discussion

Considering the description presented above we have decided to conduct the experiments, which empirically verify the influence of the ensemble learning method on the quality of the instance selection. In the experiments we examine the influence of different parameters on the compression of the training data and the accuracy of the final prediction model.

In our study, we used the Waikato Environment for Knowledge Analysis (WEKA) software V3.8.2 [14] to construct and evaluate the different models. The Weka machine learning workbench offers a general-purpose framework for automated classification, regression, clustering, and selection of features that represents common bioinformatics research data mining issues. This provides a comprehensive collection of machine learning algorithms and data pre-processing methods accompanied by graphical user interfaces for software exploration and practical comparison of various machine learning techniques on the same problem.

For the ensemble of the different used algorithms, basic hyper-parameters were fixed based on several previous tests, such as Number of Nearest neighbor for KNN Imputation $n = 1$, Number of clusters for the classification via clustering with K-means $k = 2$ and with different seed numbers (1 to 10); number of iterations of Step (c) in the approach was fixed to $t = 5$, so each time a new model is created with LOO-CV, then tested with the selected supplied test subset. As shown in Fig. 2; for each iteration of (C): correctly classified subjects are extracted from the test subset and added to the training subset then the process is repeated. For the SMOTE algorithm that based on the similarity between the available minority samples and represents the most popular and applied oversampling procedure, generates synthetic minority

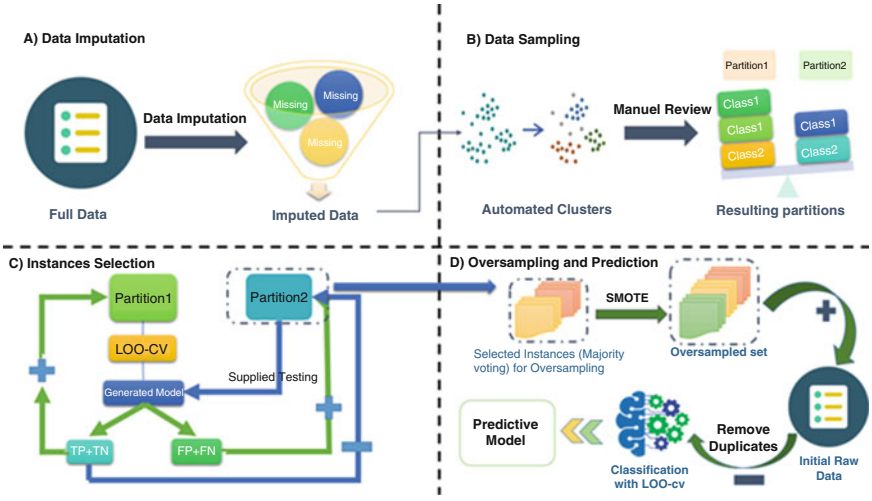

Fig. 2 Schematic representation of the proposed approach

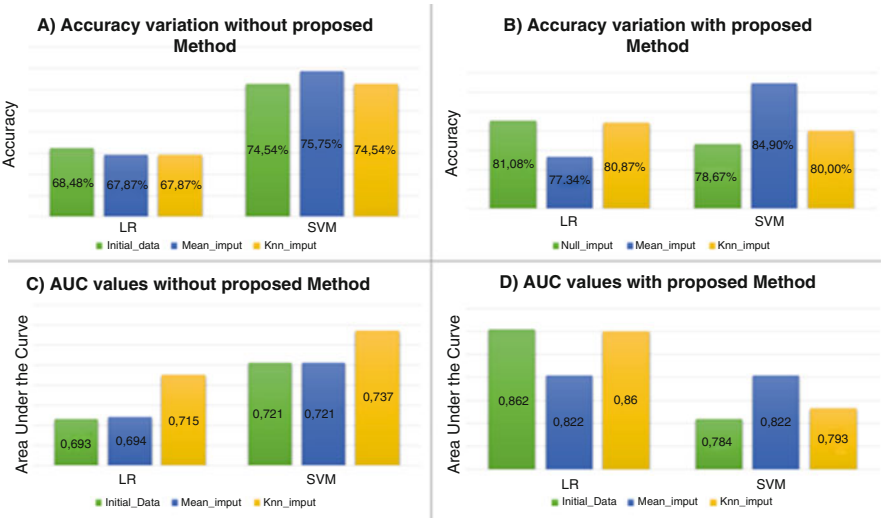

Fig. 3 Comparison the accuracy and AUC of the proposed approach achieved by the instances manipulation for different data imputation and K-means centoid Initialization

samples considering K-nearest neighbors. Thus, different nearest-neighbor values were tested nn = [1,2,3]. Logistic Regression model with ridge estimator = 1.0E-4. Support vector machine model with polynomial kernel and C = 1.0.

All of the experiments were performed on a HCC dataset composed of N = 16 patients diagnosed comprises n = 49 features [4]. Both of the tested algorithms; LR and SVM were tested independently and the results are presented using the Accuracy-Classifier plot and AUC-classifier plot to simplify the interpretation.The obtained results are presented in Figs. 3, 4, and 5. An excellent model with an AUC near to 1, means that it has a good measure of separability. Contrariwise, a poor model with an AUC near to 0 means incapability of separability between classes. When AUC equals to 0.5, it means that the model has no class separation capacity whatsoever.

Hence, in this study, several data mining and machine learning technique were applied; therefore, results were evaluated according to these applied technique. The obtained results show that our approach achieves good results coupled with imputed datasets. In several essays, data imputation didn't ameliorate the model accuracy Fig. 3a with the LR model, instead of the AUC that shows a good response to the imputation with Knn method Fig. 3c. Moreover, the SVM algorithm was sensitive for the applied imputations in terms of ACC and AUC Fig. 3a. The algorithm behaves differently when applying the proposed approach Fig. 3b, d; the accuracy increases remarkably.

Different situations were observed for the LR and SVM algorithms due to the data imputation, the centroid initialization for K-means clustering and SMOTE

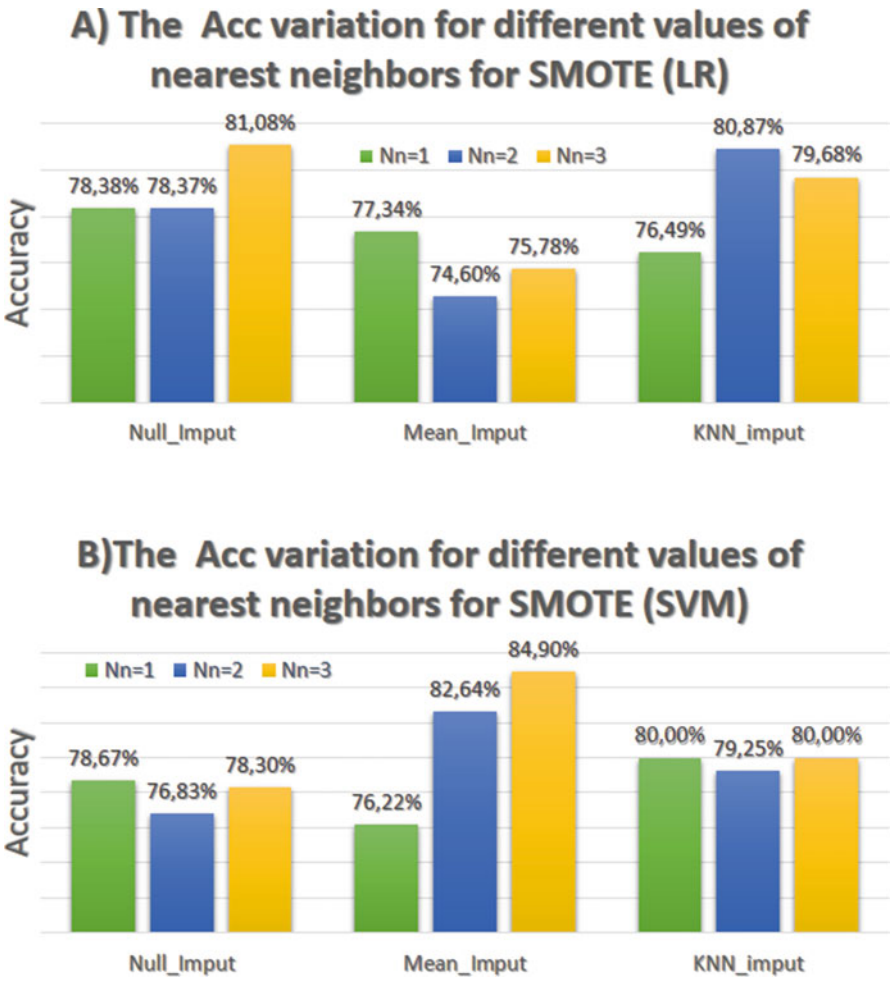

Fig. 4 Comparison the accuracy and AUC of the proposed approach achieved by the instances manipulation for different SMOTE nearest neighbor values

nearest neighbor number Fig. 4; that led to an important diversity of selected instances, hence, conducting to improve the final accuracy.

It can be seen from Fig. 4 that, our method obtained much better performance in the Accuracy and the AUC than the other methods used in [4, 5]. For different parameters of the models, we can improve the accuracy up to 84,90% without any loss of instance or features numbers. It is not possible to define a universal and optimal set of parameters. They must be chosen in each case independently, and based on several experimentation to avoid the wrong choice that cause a significant deterioration in the performance of the model. The advantages of the proposed method are as follows:

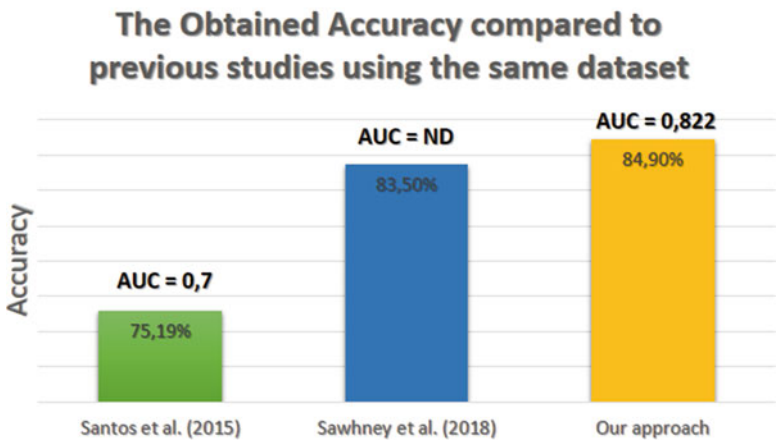

Fig. 5 The Accuracy and AUC achieved by the proposed approach other studies working on the same data set

- Obtained the highest performance (accuracy and AUC).
- Proposed method is robust and accurate as we have employed LOO-CV cross-validation twice (for the Instance selection and the prediction)
- Easy to implement and with a low computational cost.
- Instances selection was used for outliers detection and data oversampling instead of data compression.

5 Conclusion

Nowadays new possibilities open for the use of instance selection methods, in particular in limited and small data sets. These types of applications have two objectives: to improve or maintain the accuracy of the prediction model created on the selected data and to achieve the compression as high as possible. This study aimed to analyze the possibility of using ensemble learning methods to improve the efficiency of instance selection without reducing the size of data. Different data pre-processing techniques associated with learning models were used for this objective. The empirical experiments were performed with the proposed approach, which was based on instances management. The results indicate that it is possible to improve the Accuracy and the AUC while maintaining the initial data size. The approach was able to achieve preforming results compared to other approaches using the same data set.

References

1. El Houby, E.M.F.: A survey on applying machine learning techniques for management of diseases. J. Appl. Biomed. **16**, 165–74 (2018)
2. Maalel, A., Hattab, M.: 'Literature review: overview of cancer treatment and prediction approaches based on machine learning'. In: Smart Systems for E-Health, Advanced Information and Knowledge Processing, p. 324. Springer (2019). ISBN: 978-3-030-14938-3
3. Wu, C.F., Wu, Y.J., Liang, P.C., Wu, C.H., Peng, S.F., Chiu, H.W.: Disease free survival assessment by artificial neural networks for hepatocellular carcinoma patients after radiofrequency ablation. J. Formos. Med. Assoc. **116**, 765–773 (2017)
4. World Health Organization.: Hepatitis C. [online] [cit. 2017-10-15]. Available from: http://www.who.int/mediacentre/factsheets/fs164/en/ (2017)
5. Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A., Carvalho, A.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inform. **58**, 49–59 (2015)
6. Sawhney, R., Mathur, P., Shankar, R.: A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In: International Conference on Computational Science and Its Applications, pp. 438–449. Springer, Cham (2018)
7. Gamberger D., Lavrač, N.: Filtering noisy instances and outliers. In: Liu, H., Motoda, H. (eds.) Instance Selection and Construction for Data Mining. The Springer International Series in Engineering and Computer Science, vol. 608. Springer, Boston (2001)
8. Blachnik, M.: Ensembles of instance selection methods based on feature subset. Proc. Comput. Sci. **35**, 388–396 (2014)
9. Reitermanova, Z.: Data splitting. In: WDS, pp. 31–36 (2010)
10. Korjus, K., Hebart, M.N., Vicente, R.: An efficient data partitioning to improve classification performance while keeping parameters interpretable. PloS one **11**(8), e0161788 (2016)
11. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015). https://doi.org/10.1016/j.csbj.2014.11.005
12. Chang, C-C., Lin, C-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011). Article 27. https://doi.org/10.1145/1961189.1961199
13. Nebli, A., Rekik, I.: Gender differences in cortical morphological networks. Brain Imaging Behav. (2019). https://doi.org/10.1007/s11682-019-00123-6
14. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco (2011)
15. Schutte, A.E.: Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study (2017)