frontiers in Genetics

ORIGINAL RESEARCH
published: 18 September 2019
doi: 10.3389/fgene.2019.00848

Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors

Chi-Chang Chang[1,2]* *and Ssu-Han Chen*[3]*

[1] School of Medical Informatics, Chung Shan Medical University, Taichung, Taiwan, [2] IT Office, Chung Shan Medical University Hospital, Taichung, Taiwan, [3] Department of Industrial Engineering and Management, Ming Chi University of Technology, New Taipei City, Taiwan

OPEN ACCESS

Edited by:
Ping Zhang,
The Ohio State University,
United States

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Jr-Shin Li,
Washington University in St. Louis,
United States

***Correspondence:**
Chi-Chang Chang
threec@csmu.edu.tw
Ssu-Han Chen
ssuhanchen@mail.mcut.edu.tw

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *05 July 2019*
Accepted: *14 August 2019*
Published: *18 September 2019*

Citation:
Chang C-C and Chen S-H (2019) Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors. Front. Genet. 10:848. doi: 10.3389/fgene.2019.00848

Due to the high effectiveness of cancer screening and therapies, the diagnosis of second primary cancers (SPCs) has increased in women with breast cancer. The present study was conducted to develop a novel machine learning–based classification scheme for predicting the risk factors of SPCs in breast cancer survivors. The proposed scheme was based on the XGBoost classifier with the following four comparable strategies: transformation, resampling, clustering, and ensemble learning, to improve the training balanced accuracy. Results suggested that the best prediction accuracy for an empirical case is the XGBoost associated with the strategies of resampling and clustering. The experimental results showed that age, sequence of radiotherapy and surgery, surgical margins of the primary site, human epidermal growth factor, high-dose clinical target volume, and estrogen receptors are relatively more important risk factors associated with SPCs in patients with breast cancer. These risk factors should be monitored for the early detection of breast cancer. In conclusion, the proposed scheme can support the important influence of personality and clinical symptom representations in all phases of the primary treatment trajectory. Our results further suggested that adaptive machine learning techniques require the incorporation of significant variables for optimal predictions.

Keywords: second primary cancers (SPCs), breast cancer, machine learning, classification, machine learning-based classification scheme

INTRODUCTION

The effectiveness of cancer screening and therapies has resulted in an increase in the number of diagnosed second primary cancers (SPCs) throughout the world. Breast cancer is the most commonly diagnosed malignant tumor in women (Mellemkjaer et al., 2006; Kamińska et al., 2015; The Taiwan Cancer Registry, 2019a; The Taiwan Cancer Registry, 2019b). In Taiwan, breast cancer is the main type of cancer found in women. The age-adjusted incidence rates have increased from 12.07 per 100,000 women in 1979 to 73.60 per 100,000 women in 2016 (The Taiwan Cancer Registry, 2019a; The Taiwan Cancer Registry, 2019b). The five-year survival rate after breast cancer treatment has been reported to be approximately 84.97% (Huang et al., 2016). The definition of Multiple Primary Malignant Neoplasms was first published in 1932 by Warren and Gates. According to the report by Warren and Gates, both the primary and secondary tumors should be malignant with histologic confirmation, and there should be at least 2 cm of normal tissue between the two tumors. In addition, the tumors should be separated in time by at least 5 years and metastatic tumors should be excluded (Warren and Gates, 1932). In this study, we aimed to create a novel machine learning–based classification scheme for predicting the risk factors of SPCs in breast cancer survivors. Although there are several evidence-based clinical guidelines for the diagnosis and treatment of breast cancer, only a few have addressed lifelong follow-up care for breast cancer survivors. Furthermore, the treatment of breast cancer depends on the diagnostic stage, the location and size of the tumors, and tumor characteristics. Several risk factors for SPCs after treatment for breast cancer have been reported, which include environmental, smoking and alcohol use (Kolak et al., 2017), obesity (Fassier et al., 2017), cancer susceptibility genes (Yousefi et al., 2018), or previous treatments received (Markus et al., 2019).

Evidence has come from different sources, whereas, methods for synthesizing all the evidence are required. To further improve the outcomes in patients with breast cancer, physicians must identify the risk factors responsible for poor survival rates and they must develop applicable treatment strategies. Secondary cancer is due to the lack of clinical treatment strategies as well as the absence of risk factor identification to prevent its occurrence.

Many studies have been conducted using statistical methods for cancer classification and predictions (Liu et al., 2011; Khormuji and Bazrafkan, 2016; Xie et al., 2016). However, these statistical models require the establishment of formidable model assumptions during the model construction process. When these modeling assumptions are violated, it becomes difficult to achieve the desired results. Unlike models for statistical disease prediction, cancer prediction models that are based on machine learning techniques do not require powerful model assumptions and *a priori* assumptions concerning the properties of the data. They can, however, capture delicate underlying patterns and relationships contained in empirical data and they provide promising cancer prediction results (Tseng et al., 2014; Kourou et al., 2015; Tseng et al., 2017; El Houby, 2018).

Machine learning–based cancer classification models have been used in many reports in the literature to predict breast cancer recurrence (Yu et al., 2014; Ye et al., 2018; Vural et al., 2016). However, to the best of our knowledge, no reported studies have proposed machine learning–based classification schemes for SPC classifications.

MATERIALS AND METHODS

We used machine learning techniques to develop a novel classification scheme, which included transformation of data, clustering, resampling, and ensemble learning (TCRE) to predict SPCs in women to have had breast cancer. In the proposed model, we first divided the original dataset into training data and testing data using specific percentages. **Figure 1** shows the flowchart for this scheme.

In the proposed TCRE scheme (**Figure 1**), the original dataset was divided into the training data and testing data with specific percentages. Both the training and testing datasets were arranged into a clear form in which the columns represent the features, and labels and rows represent the cases. Subsequently, a series of procedures were conducted using the following steps:

Step 1 includes transformation to determine better feature representation. Principal component analysis (PCA) was used to transform the original feature space into a lower dimensional space in which each dimension could be regarded as a base that explains the variability of the data best, which is similar to a noise removal process. PCA has been empirically proven in the literature to be able to improve classification results (Trivedi et al., 2015; Ikram and Cherukuri, 2016; Nasution et al., 2018). The classifier of ensemble learning, which is described below, used the gradient descent idea in an iterative manner to identify parameters with a local minimum. The classifier always disapproves the problem curse of dimensionality and, therefore, maintaining less and insignificant data may improve the convergence speed and the quality of the classification results.

Step 2 includes clustering to group cases that are similar in advance. Previous studies have indicated that performing clustering before classification may be beneficial because new grouping information is assigned in a dummy fashion relative to the original dataset (Alapati and Sindhu, 2016; Sekula et al., 2017). The k-means or k-modes algorithm has been

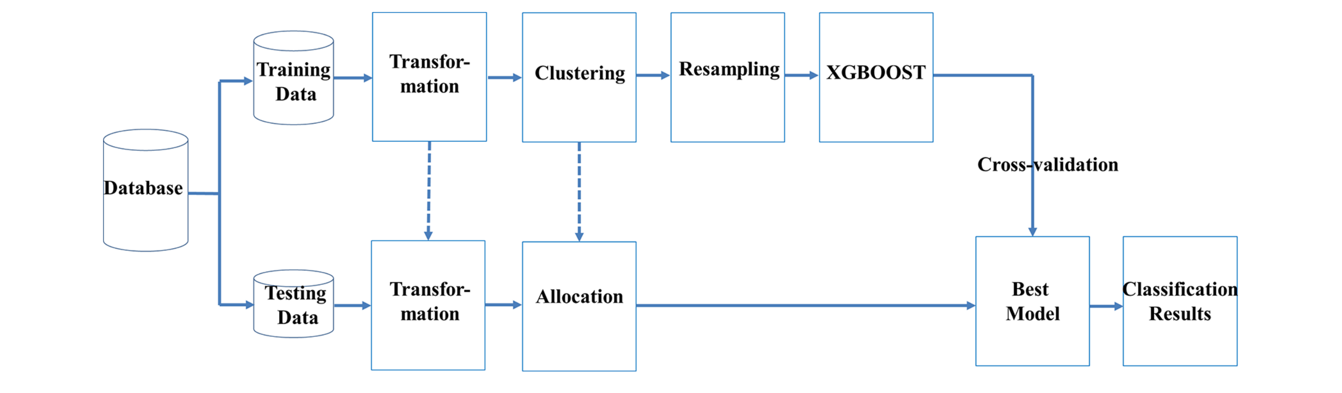

FIGURE 1 | Flowchart of the proposed TCRE scheme.

used for clustering training data, in which the optimal clusters number is determined by internal validation measures (Chawla et al., 2004) and rank aggregation of stability (Huang, 1998).

Step 3 includes resampling to alleviate the class imbalance issue. The SPBC datasets generally have class imbalance problems because they often comprise a much higher number of breast cancer patients without SPBCs, even though only a small percentage of patients have an SPBC. When the datasets have imbalanced classes, the classifiers struggle for accuracy with imbalanced data because they are biased toward the majority class. Worse yet, the classifiers may predict everything to belong to the majority and the minority is therefore ignored to pursue a high, but pseudo, accuracy rate (Wang and Yao, 2012; Wang and Yao, 2013; Chen and Guestrin, 2016). In this study, we focus on prediction improvements through resampling method by applying SMOTE (Synthetic Minority Oversampling TEchnique) to conduct data resampling. SMOTE builds upon two methods by up-sampling the minority class and down-sampling the majority class (Fernández et al., 2018). In addition, we applied oversampling technique and try to preprocess imbalanced data before we feed them into a classifier. The main motivation behind the need to detecting the majority class and less sensitive to the minority class.

Step 4 uses ensemble learning to construct an effective classifier that can classify patients diagnosed with SPCs accurately. In step 4, the eXtreme gradient boosting (XGBoost), which was proposed by Chen and Guestrin (2016), is built based on the principles of gradient boosting trees. Trees can be efficiently constructed, and computations can be operated in parallel. The XGBoost was used in this study because it is an effective ensemble learning algorithm that can be used for various medical issues (Schmidhuber et al., 2018; Sun et al., 2018). Other reasons for choosing the XGBoost include the presence of several ordered or categorical variables in the dataset, there is no requirement for a data distribution assumption, and tree-based methods often perform well on imbalanced datasets.

When using XGBoost in the proposed TCRE scheme, the primary question was how to tune the hyperparameters of this classifier during the training process to produce a model with a performance that is relatively better. As XGBoost is a flexible classifier that provides numerous hyperparameters, such as the eta, maximum depth of a tree, number of rounds for boosting, gamma, and the subsample ratio of columns (Chen and Guestrin, 2016). When constructing each tree of XGBoost, the sum of an instance weight in a child and the subsample must be minimized. However, it is nearly impossible to manually choose a good set of hyperparameter combinations. The commonly used methods to resolve this problem combine the processes of k-fold cross-validation, a random search, and metric evaluation. **Figure 2** depicts the proposed procedure for identifying the best hyperparameter set for XGBoost.

Figure 2 shows the random search scheme that implements a randomized search over the hyperparameters for *m* times in which each value of the setting is derived from a uniform distribution over all possible values of the hyperparameters. Then, for each set of hyperparameters, the training data is randomly divided into *k* equal-sized folds. Of the *k* folds, the k-1 folds are used as the "really" training data for training the model, and the remaining single fold is considered as the validation data for validating the performance of the corresponding hyperparameters. This process is repeated *k* times, and then the corresponding evaluation metric is calculated and subsequently averaged to produce an average

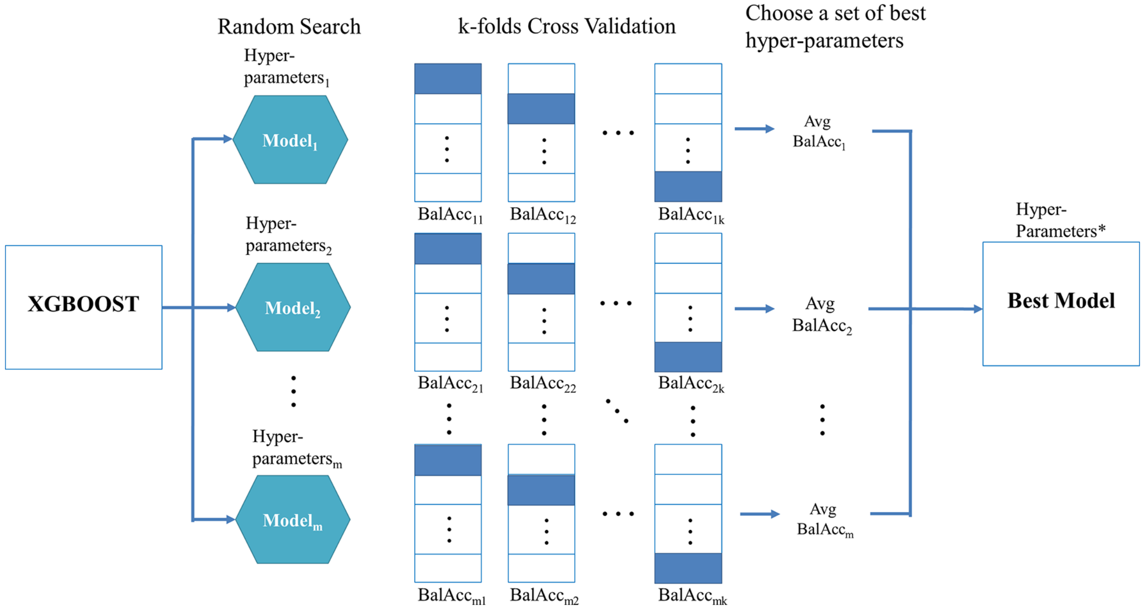

FIGURE 2 | The procedure for determining the best hyperparameter set for the used XGBoost.

value for each set of hyperparameters. The metrics, such as AUC, kappa, or balanced accuracy, are suggested in the class imbalanced dataset rather than sensitivity, specificity, or accuracy because the former set of metrics simultaneously takes the performance of each class into consideration. In particular, the balanced accuracy was adopted in this study, which considered the average values of sensitivity and specificity.

In the testing phase, as shown in **Figure 3**, each sample is transformed to PCA space based on the weight matrix and the mean vector of the training data, and then the sample is allocated to its nearest clustering center. Only the features of the preprocessed testing data were fed into the best model to obtain the corresponding prediction responses. Lastly, the prediction responses were compared with the corresponding labels in the testing data to generate a confusion matrix. Furthermore, the information about variable importance was also extracted based on the training information from the best models, which was then used to identify important risk factors for breast cancer survivors.

RESULTS AND DISCUSSION

The medical records and the corresponding pathologic status provided by Chung Shan Medical University Hospital, Jen-Ai Hospital, and Far Eastern Memorial Hospital Tumor Registry were used for training and for evaluating the proposed methodology. The 23 predictor variables analyzed in this paper are considered to be associated with the risk factors for secondary cancer. On the basis of the comments by the expert committee and the properties of the data, the predictor variables are as follows: 1) age, 2) primary site, 3) histology, 4) behavior code, 5) differentiation, 6) tumor size, 7) pathologic stage, 8) surgical margin of the primary site, 9) surgery, 10) radiotherapy (RT), 11) RT surgery, 12) sequence of local regional therapy and systemic therapy, 13) sequence of radiotherapy and surgery, 14) dose to clinical target volume (CTV) high, 15) number to CTV high, 16) dose to CTV low, 17) number to CTV low, 18) body mass index (BMI), 19) smoking, 20) drinking, 21) human epidermal growth factor (HER2), 22) estrogen receptors (ERs), and 23) progesterone receptors.

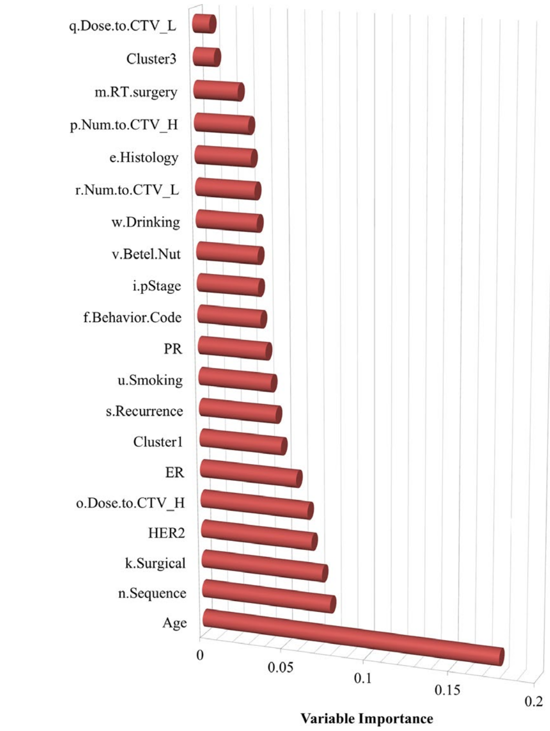

FIGURE 3 | The ranked variable importance for breast cancer survivors.

A total of 2,964 patients diagnosed with breast cancer in three hospitals between 2010 and 2016 were included in the study. The study population involved 185 SPC cases with breast cancer. Our dataset suffered from the class imbalance problem because the total number in the class of breast cancer survivors was far less than the total number of another class of breast cancer survivors without SPCs. In addition, the dataset was randomly divided by 60% and 40% with respect to the training and testing datasets, respectively. The majority of existing studies directly use machine learning methods for cancer classification without using transformation, clustering, and resampling to deal with preprocessing and class imbalance issues (Tseng et al., 2017). Next, the built classifiers tended to predict only the majority class data, which results in a high misclassification rate of minority classes when compared with the majority class. The classification result after analysis of the data by directly using XGBoost without using transformation, clustering, and resampling (known as the single XGBoost method, which is used as a benchmark method) is shown in the first row of **Table 1**. It can be observed that the single XGBoost method provided a very high testing accuracy of >94%. However, the testing balanced accuracy was only 49.95%, which implies that all cases were classified as the majority class. Thus, the classifier single XGBoost learned nothing.

The Sensitivity Analysis and the Corresponding Classification Results

The proposed TCRE scheme includes the use of PCA, resampling, clustering, and XGBoost. A combination of these parameters was adopted during the training and testing processes to evaluate the performance of each preprocessing method. Such a combination implies using or not using PCA transformation, applying or not applying resampling, performing or not performing clustering before classification. However, currently, we have no idea about whether each preprocessing method was adopted in association with the model in this dataset. In addition, no specified preprocessing method combination has been confirmed to be the best; it depends on the available dataset. Therefore, a sensitivity analysis must be conducted while users are training a model. In addition, the accuracy was (TP+TN)/(P+N), where TP is a true positive, TN is a true negative, P is the number of real positive cases in the data and N is the number of real negative cases in the data. The balanced accuracy was used to deal with imbalanced datasets. The balanced accuracy is defended as (TP/P+TN/N)/2. **Table 1** presents the results of the sensitivity analysis and the corresponding classification results of the proposed TCRE scheme. **Supplementary Materials** show the detailed information as determined by the testing balanced accuracy and validation balanced accuracy within the proposed TCRE scheme.

TABLE 1 | Results of the sensitivity analysis and the corresponding classification results of the proposed TCRE scheme.

Transformation	Resampling	Clustering	Training accuracy	Training balanced accuracy	Testing accuracy	Testing balanced accuracy
0	0	0	0.9447	0.5104	0.9432	0.4995
0	0	1	0.9447	0.5104	0.9387	0.4971
0	1	0	0.7195	0.7604	0.6853	0.5830
0	1	1	0.7075	0.7743	0.6889	0.6000
1	0	0	0.9526	0.5803	0.9378	0.5042
1	0	1	0.9471	0.5775	0.9342	0.5023
1	1	0	0.6691	0.6629	0.6583	0.5307
1	1	1	0.6348	0.6700	0.6348	0.5638

In the first three columns of **Table 1**, the numeral 1 represents the corresponding preprocessing methods that are used, whereas, the zeros represent the methods that are not activated in this study. In the subsequent columns, the accuracy was calculated as the proportion of cases that were correctly classified, while the balanced accuracy was defined as the average value of the proportion correct in each individual class. The term training or testing represents the abovementioned metrics generated during either the training stage or the testing stage. We lastly aimed to maximize the training balanced accuracy based on the analysis. The process suggests that both resampling and clustering techniques must be adopted in the subsequent testing stage. On the basis of the optimal model selected throughout the training process described above, the rate of SPCs in women with breast cancer in the testing data was also approximately 6.2%. As shown in **Table 1**, our results show that the best scheme was the XGBoost associated with the strategies of resampling and clustering. In addition, the performance of the testing balanced accuracy and training balanced accuracy were increased by 10.05% and 26.39%, compared to previous the baseline XGBoost.

Identification of Important Risk Factors

Variable importance was also assessed, as shown in **Figure 3**, which indicates the features that are more influential on patients with SPBC. It determined that age, the sequence of radiotherapy and surgery, surgical margins of the primary site, HER2, dose to CTV high, and ER are relatively more important risk factors associated with SPCs.

CONCLUSIONS

On the basis of the evidence obtained in this study, it can be concluded that the positive correlation between breast cancer and SPCs is not an accidental result. Breast cancer is the most common female cancer throughout the world. Although studies of breast cancer survivors dominate the survivor literature, few prospective randomized controlled trials have intervened in breast cancer survivors. We urge caution regarding the prevention and treatment of breast cancer survivors. Risks of long-term and late-stage effects after breast cancer treatment are associated with several factors. The results of this study suggest that age, sequence of radiotherapy and surgery, surgical margins of the primary site, HER2, dose to CTV high, and ER, when appropriate, should be recommended for patients with breast cancer. Radiation, chemotherapy, and hormone/endocrine therapy with aromatase inhibitors are especially associated with an increased risk of developing SPCs in patients with breast cancer. There is sufficient evidence indicating that obesity is a risk factor for SPC development and other problems. To determine whether women with breast cancer are genetically susceptible or at high risk of developing SPCs that may affect other family members, it is necessary to collect their detailed medical history, including key risk factors, and the family history of their parents. Long-term follow-up of patients with breast cancer is important for documenting the risks and patterns of SPCs, and knowledge about these aspects will influence surveillance and prevention strategies in the future.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

C-CC initially drafted the manuscript and collected the features, analyzed the experiments and revised the paper. S-HC did part of the codes work and the experiments. All authors designed the work, read and approved the final manuscript and are agree to be accountable for all aspects of the work.

FUNDING

This work was supported by the MOST 106-2218-E-040 -001 -MY2.

ACKNOWLEDGMENTS

We thank the reviewers for their great comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00848/full#supplementary-material

REFERENCES

Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., and Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Comput. Surv.* 49, 52. doi: 10.1145/2988544

Alapati, Y. K., and Sindhu, K. (2016). Combining clustering with classification: a technique to improve classification accuracy. *Lung Cancer* 32, 3.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explor. Newsl.* 6, 1–6. doi: 10.1145/1007730.1007733

Chen, T., and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 785-794, ACM doi: 10.1145/2939672.2939785

El Houby, E. M. (2018). Framework of computer aided diagnosis systems for cancer classification based on medical images. *J. Med. Syst.* 42, 157. doi: 10.1007/s10916-018-1010-x

Fassier, P., Zelek., L., Bachmann., P., Touillaud., M., Druesne-Pecollo., N., Partula., V., et al. (2017). Sociodemographic and economic factors are associated with weight gain between before and after cancer diagnosis: results from the prospective population-based NutriNet-Santé cohort. *Oncotarget* 8, 54640–54653. doi: 10.18632/oncotarget.17676

Fernández, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* 61, 863–905. doi: 10.1613/jair.1.11192

Huang, H. H., Liu, X. Y., and Liang, Y. (2016). Feature selection and cancer classification *via* sparse logistic regression with the hybrid L1/2+2 regularization. *PLoS One* 11, e0149675. doi: 10.1371/journal.pone.0149675

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* 2, 283–304. doi: 10.1023/A:1009769707641

Ikram, S. T., and Cherukuri, A. K. (2016). Improving accuracy of intrusion detection model using PCA and optimized SVM. *J. Comput. Sci. Tech.* 24, 133–148. doi: 10.20532/cit.2016.1002701

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., and Starosławska, E. (2015). Breast cancer risk factors. *Prz Menopauzalny* 14, 196–202. doi: 10.5114/pm.2015.54346

Khormuji, M. K., and Bazrafkan, M. (2016). A novel sparse coding algorithm for classification of tumors based on gene expression data. *Med. Biol. Eng. Comput.* 54, 869–876. doi: 10.1007/s11517-015-1382-8

Kolak, A., Kamińska, M., Sygit, K., Budny, A., Surdyka, D., Kukielka-Budny, B., et al. (2017). Primary and secondary prevention of breast cancer. *Ann. Agric. Environ. Med.* 24, 549–553. doi: 10.26444/aaem/75943

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Liu, C., Pan, C., Shen, J., Wang, H., and Yong, L. (2011). MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of colorectal cancer. *Int. J. Med. Sci.* 8, 39–47. doi: 10.7150/ijms.8.39

Markus, E., Cristoforo, S., Pavel, K., Alexander, U., Elena, S., Denise, G., et al. (2019). Long-term health risk after breast-cancer radiotherapy: overview of PASSOS methodology and software. *Radiat. Prot. Dosimetry* 183 (1-2), 259–263. doi: 10.1093/rpd/ncy219

Mellemkjaer, L., Friis, S., Olsen, J. H., Scélo, G., Hemminki, K., Tracey, E., et al. (2006). Risk of second cancer among women with breast cancer. *Int. J. Cancer.* 118, 2285–2295. doi: 10.1002/ijc.21651

Nasution, M. Z. F., Sitompul, O. S., and Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4. 5 classification. *J. Phys. Conf.* 978, 012058. doi: 10.1088/1742-6596/978/1/012058

Schmidhuber, J., Sur, P., Fay, K., Huntley, B., Salama, J., Lee, A., et al. (2018). The Global Nutrient Database: availability of macronutrients and micronutrients in 195 countries from 1980 to 2013. *Lancet Planet. Health* 2, e353–e368. doi: 10.1016/S2542-5196(18)30170-0

Sekula, M., Datta, S., and Datta, S. (2017). optCluster: An R package for determining the optimal clustering algorithm. *Bioinformation.* 13, 101–103. doi: 10.6026/97320630013101

Shimoda, A., Ichikawa, D., and Oyama, H. (2018). Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Comput. Methods Programs Biomed.* 163, 39–46. doi: 10.1016/j. cmpb.2018.05.032

Sun, B., Lam, D., Yang, D., Grantham, K., Zhang, T., Mutic, S., et al. (2018). A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Med. Phys.* 45, 2243–2251. doi: 10.1002/mp.12842

The Taiwan Cancer Registry (2019a). The 5-year survival rate after breast cancer. 2019.07.01[Access date], url: http://tcr.cph.ntu.edu.tw/uploadimages/Survival_101_105.pdf.

The Taiwan Cancer Registry (2019b). The age-adjusted incidence rates: 1979-2016. 2019.07.01[Access date], url: http://tcr.cph.ntu.edu.tw/main.php?Page=A5B2.

Trivedi, S., Pardos, Z. A., and Heffernan, N. T. (2015). The utility of clustering in prediction tasks. *arXiv preprint arXiv* 2015, 1509.06163.

Tseng, C. J., Lu, C. J., Chang, C. C., and Chen, G. D. (2014). Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput. Appl.* 24, 1311–1316. doi: 10.1007/s00521-013-1359-1

Tseng, C. J., Lu, C. J., Chang, C. C., Chen, G. D., and Cheewakriangkrai, C. (2017). Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif. Intell. Med.* 78, 47–54. doi: 10.1016/j.artmed.2017.06.003

Vural, S., Wang, X., and Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst. Biol.* 10, 62. doi: 10.1186/s12918-016-0306-z

Wang, S., and Yao, X. (2012). Multiclass imbalance problems: analysis and potential solutions. *IEEE Trans. Syst. Man Cybern. B Cybern.* 42, 1119–1130. doi: 10.1109/TSMCB.2012.2187280

Wang, S., and Yao, X. (2013). Using class imbalance learning for software defect prediction. *IEEE T. Reliab.* 62, 434–443. doi: 10.1109/TR.2013.2259203

Warren, S., and Gates, O. (1932). Multiple malignant tumors. A survey of the literature and statistical study. *Am. J. Cancer.* 16, 1358–1414.

Xie, H., Li, J., Zhang, Q., and Wang, Y. (2016). Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Comput. Biol. Chem.* 65, 165–172. doi: 10.1016/j.compbiolchem.2016.09.010

Ye, L., Lee, T. S., and Chi, R. (2018). A hybrid machine learning scheme to analyze the risk factors of breast cancer outcome in patients with diabetes mellitus. *J. Univers. Comput. Sci.* 24, 665–681. doi: 10.3217/jucs-024-06-0665

Yousefi, M., Nosrati, R., Salmaninejad, A., Dehghani, S., Shahryari, A., and Saberi, A. (2018). Organ-specific metastasis of breast cancer: molecular and cellular mechanisms underlying lung metastasis. *Cell. Oncol. (Dordr).* 41, 123–140. doi: 10.1007/s13402-018-0376-6

Yu, X., Chum, P., and Sim, K. B. (2014). Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system. *Optik (Stuttg)* 125, 1498–1502. doi: 10.1016/j.ijleo.2013.09.013

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.