


Article

On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer Prediction

Min-Wei Huang ^{1,2,†}, Chien-Hung Chiu ^{3,†}, Chih-Fong Tsai ⁴ and Wei-Chao Lin ^{3,5,*} 

¹ Department of Physical Therapy and Graduate Institute of Rehabilitation Science, China Medical University, Taichung 406040, Taiwan; hminwei@gmail.com

² Department of Psychiatry, Chiayi Branch, Taichung Veterans General Hospital, Chiayi 60090, Taiwan

³ Department of Thoracic Surgery, Chang Gung Memorial Hospital, Linkou 333423, Taiwan; b9102067@cgmh.org.tw

⁴ Department of Information Management, National Central University, Taoyuan 320317, Taiwan; cftsai@mgt.ncu.edu.tw

⁵ Department of Information Management, Chang Gung University, Taoyuan 33302, Taiwan

* Correspondence: viclin@gap.cgu.edu.tw

† These authors contributed equally.

Abstract: Breast cancer prediction datasets are usually class imbalanced, where the number of data samples in the malignant and benign patient classes are significantly different. Over-sampling techniques can be used to re-balance the datasets to construct more effective prediction models. Moreover, some related studies have considered feature selection to remove irrelevant features from the datasets for further performance improvement. However, since the order of combining feature selection and over-sampling can result in different training sets to construct the prediction model, it is unknown which order performs better. In this paper, the information gain (IG) and genetic algorithm (GA) feature selection methods and the synthetic minority over-sampling technique (SMOTE) are used for different combinations. The experimental results based on two breast cancer datasets show that the combination of feature selection and over-sampling outperform the single usage of either feature selection and over-sampling for the highly class imbalanced datasets. In particular, performing IG first and SMOTE second is the better choice. For other datasets with a small class imbalance ratio and a smaller number of features, performing SMOTE is enough to construct an effective prediction model.

Keywords: breast cancer; data mining; machine learning; feature selection; over-sampling; class imbalance



Citation: Huang, M.-W.; Chiu, C.-H.; Tsai, C.-F.; Lin, W.-C. On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer Prediction. *Appl. Sci.* **2021**, *11*, 6574. <https://doi.org/10.3390/app11146574>

Academic Editors: Stefano Silvestri and Francesco Gargiulo

Received: 16 June 2021

Accepted: 15 July 2021

Published: 17 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Breast cancer, which is cancer that develops from breast tissue, is one of the important problems in the medical domain. It is the second most severe cancer among all of the cancers that have already been discovered. Some factors have been found to cause breast cancer, such as obesity, a lack of physical exercise, alcoholism, hormone replacement therapy during menopause, ionizing radiation, a family history of breast cancer, etc. [1]. In practice, many medical institutes have paid much attention to the early detection of breast cancer.

In related literatures, many data mining and machine learning techniques have been used to develop various kinds of breast cancer prediction models. Among them, some focus on the improvement of learning models and some focus on data pre-processing steps. For example, convolutional neural networks (CNN), as one representative of a deep learning technique, were modified to improve their prediction performance [2,3]. On the other hand, some studies focus on feature selection for filtering out irrelevant features from a given dataset for the construction of more effective classifiers [4,5] and data sampling

for re-balancing class imbalanced datasets in order to decrease the effect of skewed class distribution in the learning process [6,7].

For related works of feature selection, Sasikala et al. [8] propose a novel feature selection method based on the genetic algorithm to select a gene subset from high dimensional gene data, which causes different classifiers perform better than the ones without feature selection. In [9], a genetic algorithm is used for feature selection, where the selected subset is used to construct different classifiers for performance comparisons. On the other hand, Jiang and Jin [10] use a gradient boosting decision tree with Bayesian optimization to remove the irrelevant and redundant features from gene expression data. Raj et al. [11] compare several feature selection methods to determine the best one to combine with the random forest classifier.

For related works on class imbalance learning, Zhang et al. [12] propose a clustering-based under-sampling method to select informative samples from the clusters identified in the majority and minority classes, and the decision tree based on this boosting technique is employed for the prediction model. In [13], eighteen different under- and over-sampling methods are used to balance related class imbalanced cancer datasets, in which the over-sampling methods perform better than the under-sampling ones. Cai et al. [14] apply the synthetic minority over-sampling technique (SMOTE) to balance the training dataset and employ the stacking ensemble method to combine multiple classifiers, which achieved better performance than conventional methods. Rani et al. [15] investigated the effect of performing SMOTE on five different classifiers to determine the best one for breast cancer prediction.

According to Fernandez et al. [16], SMOTE over-sampling can benefit from the use of feature selection, where feature selection is performed over the class imbalanced dataset to select a subset feature of it, and then the reduced dataset is over-sampled to make it contain the same size of the data samples as in the majority and minority classes. Recently, Solanki et al. [17] propose the contrary procedure that SMOTE be performed first to re-balance the breast cancer dataset, and then wrapper-based feature selection methods can be applied to reduce the feature dimensions.

However, to the best of our knowledge there is not any study examining the performances of both procedures to combine feature selection and over-sampling for breast cancer prediction. Therefore, the research objective of this paper is to compare these two combination orders with two baselines by employing feature selection and over-sampling individually. Particularly, filter and wrapper-based feature selection methods are combined with SMOTE for performance comparison. In addition, one small- and one large-scale breast cancer datasets are used in order to understand the performance of different approaches.

The contribution of this paper is two-fold. First, the procedures of combining the feature selection and over-sampling steps are compared in terms of breast cancer prediction, which has never been done before. Second, the best combination procedure and combined algorithms that will be identified in this paper can be used as one the representative baseline methods for future research.

The rest of this paper is organized as follows. Section 2 overviews related literature on feature selection and over-sampling. Section 3 describes the two different combination procedures and the experimental setup. Section 4 presents the experimental results, and Section 5 concludes the paper.

2. Literature Review

2.1. Feature Selection

Feature selection is an important data pre-processing step in data mining and knowledge discovery from databases. It focuses on selecting representative features from a given training set, which have higher discriminative power to make classifiers better able to distinguish between different classes. Moreover, another advantage of feature selection is

to reduce feature dimensionality, which lowers the computational complexity during the classifier training stage [5,18].

In general, feature selection algorithms are composed of four basic steps, which are a generation procedure to generate the candidate feature subset, an evaluation function to evaluate the effectiveness of the feature subset, a stopping criterion to determine when to stop the previous steps, and a validation procedure to examine whether the feature subset is valid [19].

Existing feature selection algorithms can be divided into filter, wrapper, and embedded methods depending on how they combine the feature selection search with the construction of the classifiers. In filter methods, the relevance of features such as distance, consistency, dependency, information, and correlation are assessed. That is, the feature relevance score is calculated, in which low-scoring features are removed. Some representative methods include relief, the Fisher score, and information gain.

In wrapper methods, a specific classification algorithm is used to determine the quality of different subsets of features. Since the space of feature subsets can grow exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset. Therefore, wrapper methods are very computationally intensive, especially when the construction of the chosen classifier requires a high computational cost. One representative wrapper method is the genetic algorithm.

In embedded methods, feature selection is incorporated as part of the classifier training process. That is, the feature selection method is embedded in the modeling algorithm, where the classifier is used to evaluate the quality of the selected subset of features. Embedded methods have the advantage of including interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods. One representative wrapper method is the decision tree classifier.

2.2. Over-Sampling

In practice, the class imbalanced dataset problem usually occurs since the number of data samples in one class are significantly different from those of the other one; say the imbalance ratio is 1:100. For the example of breast cancer datasets, they do not usually contain both the malignant and benign patient classes, denoted as the minority and majority classes, respectively. Without dealing with the class imbalance problem, most machine learning models aim at maximizing the accuracy of its classification rule by ignoring the minority class examples, with the classification of all testing examples being organized into the majority class [6].

In general, there are three types of solutions to the class imbalance problem, which are algorithm level, data level, and cost-sensitivity methods. Among them, the data level methods based on data sampling techniques are usually considered first since they are used independently of the classifier [6]. Data sampling techniques focus on re-balancing the given training set. Particularly, under- and over-sampling techniques have been used, in which the former is for reducing the size of the majority class, whereas the latter is used for enlarging the size of the minority class. Among them, the synthetic minority over-sampling technique (SMOTE) is one representative method, which has been used as the baseline in many related studies [16].

The aim of SMOTE is to produce new synthetic examples for the minority class. For example, a minority class instance i is selected as the basis to create new synthetic data. According to a specific distance metric, usually the Euclidean distance, the number of the neighbors nearest to i are chosen from the training set, e.g., i_1 , i_2 , and i_3 . Next, a randomized interpolation is conducted to obtain new synthetic data, i.e., s_1 , s_2 , and s_3 .

3. Research Methodology

3.1. Two Combination Orders for Feature Selection and Over-Sampling

In this paper, two orders of combining the feature selection and over-sampling steps are compared by being given a training set, denoted as TR , which is composed of M and N

majority and minority class data samples, respectively, and each data sample is represented by k dimensional features. For the first order, i.e., performing feature selection first and over-sampling second, a chosen feature selection algorithm is employed to select some representative features from the TR . As a result, a reduced feature subset of TR is produced, denoted as $TR_{reduced}$, where each data sample is represented by o dimensional features ($k > o$). Next, the over-sampling algorithm is used to generate $M-N$ synthetic data samples for the minority class, leading to a balanced training set, denoted as $TR_{reduced_balanced}$, which is composed of $2M$ data samples. That is, the number of data samples in the majority and minority classes are the same.

On the other hand, for the second combination order, the over-sampling algorithm is used first to re-balance the training set, i.e., TR , which results in a balanced training set, denoted as $TR_{balanced}$. $TR_{balanced}$ is composed of $2M$ data samples, and each data sample is represented by k dimensional features. Next, the chosen feature selection algorithm is performed over $TR_{balanced}$, leading to a reduced feature subset of $TR_{balanced}$, denoted as $TR_{balanced_reduced}$. In $TR_{balanced_reduced}$, each data sample is represented by p dimensional features ($k > p$). Note that the number of features in $TR_{reduced_balanced}$ by the first combination order and $TR_{balanced_reduced}$ by the second combination order are not necessarily the same, i.e., $o \neq p$.

Therefore, the performances of the classifiers trained by $TR_{reduced_balanced}$ and $TR_{balanced_reduced}$ can be compared individually based on the same testing set. Moreover, other classifiers trained by $TR_{reduced}$ through performing feature selection alone and $TR_{balanced}$ through performing over-sampling alone are regarded as the baseline approaches for further performance comparison.

3.2. Experimental Setup

3.2.1. Datasets

In order to examine the performances of both orders of combining feature selection and over-sampling, two related breast cancer datasets are considered. The first one is based on the KDD Cup 2008 breast cancer dataset (<https://www.kdd.org/kdd-cup/view/kdd-cup-2008> (accessed on 15 February 2021)), which contains 102294 data samples, and each data sample is represented by 117 different image features, which are extracted from 4 X-ray images per patient. Particularly, the class imbalance ratio is 163.2.

The second dataset is based on the Breast Cancer Wisconsin Dataset downloaded from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29> (accessed on 15 February 2021)). It is composed of 699 data samples, in which each data sample is represented by 10 features including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. In addition, the class imbalance ratio is 1.86.

To train and test the classifier, the 5-fold cross validation method is used to divide each dataset into 80% and 20% training and testing sets. This means that every subset will be trained and tested five times, and the average prediction accuracy can be obtained consequently. In other words, each patient data will be used as the training and testing data example. In addition, the class imbalance ratio of the training set in each fold is controlled to be the same as the original dataset.

3.2.2. The Feature Selection and Over-Sampling Methods

In this paper, the information gain (IG) as the filter method and the genetic algorithm (GA) as the wrapper method are used for feature selection. Particularly, these two methods have been used in many research problems, including text classification [20], gene expression microarray analysis [21], intrusion detection [22], financial distress prediction [23], software defect prediction [24], etc.

IG evaluates the gain of each variable in the context of the target variable, which is based on calculating the reduction in entropy. That is, the feature ranking stage focuses on

ranking the subsets of features by high information gain entropy in decreasing order. In GA, an initial set of candidate solutions (i.e., individuals) are created and their corresponding fitness values are calculated for the later cross-over and mutation steps. Specifically, the individuals are subsets of predictors, and the fitness values are measures of the model performance.

Analyses were performed using the WEKA data mining software package. Most related parameters are based on its default values, except for the genetic algorithm, where the population size, crossover rate, and mutation rate were set as 50, 0.8, and 0.01, respectively [25].

On the other hand, the over-sampling method is based on SMOTE. It has been widely used as a baseline over-sampling method for breast cancer datasets [14–17]. The percentage of synthetic instances was set to make the two datasets become balanced datasets where the malignant and benign classes contain the same numbers of data samples. Other related parameters were based on the default values of WEKA.

3.2.3. The Classifier Design

After the original training set TR was pre-processed by different approaches, i.e., $TR_{reduced_balanced}$, $TR_{balanced_reduced}$, $TR_{balanced}$, and $TR_{reduced}$, they were used to train the support vector machine (SVM) classifier for performance comparisons. In related literature, SVM has been widely used as the baseline classifier for breast cancer prediction [26–29].

The implementation of SVM was based on the RBF kernel function, and its related parameters were based on the default values of WEKA.

4. Experimental Results

4.1. The KDD Cup 2008 Breast Cancer Dataset

Figure 1 shows the AUC (area under the ROC curve) rates of different approaches. In addition, Figure 2 shows the type I errors of the different approaches, which represent the error of miss-classifying the malignant cases into the benign class. Note that IG+SMOTE and GA+SMOTE mean the combination order of performing feature selection first and over-sampling second, whereas SMOTE+IG and SMOTE+GA represent the opposite combination order. In addition, the baseline represents using the original training set without performing any feature selection or over-sampling steps to train the SVM classifier.

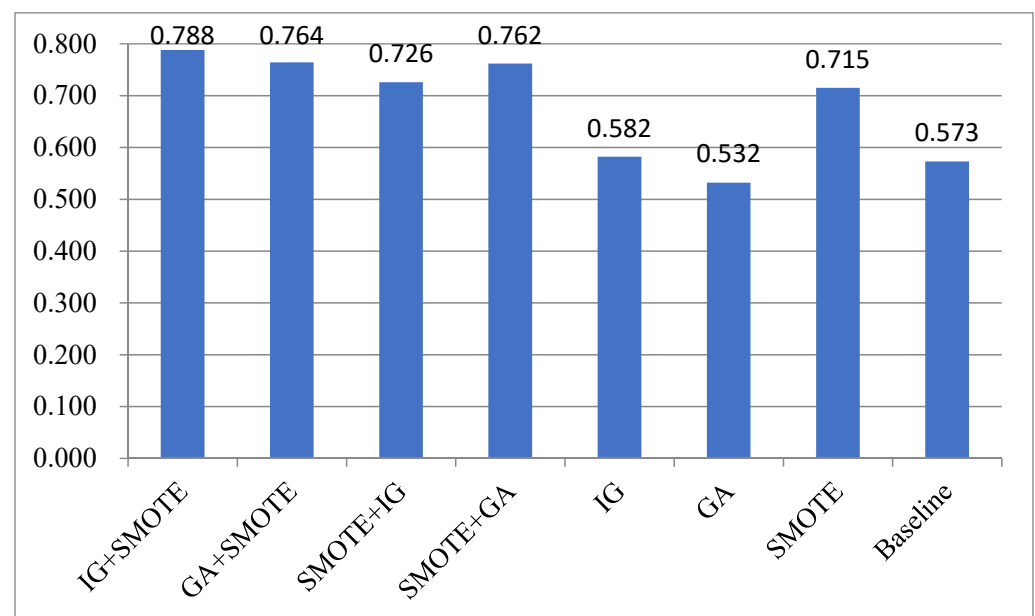


Figure 1. AUC rates of different approaches over the KDD Cup 2008 breast cancer dataset.

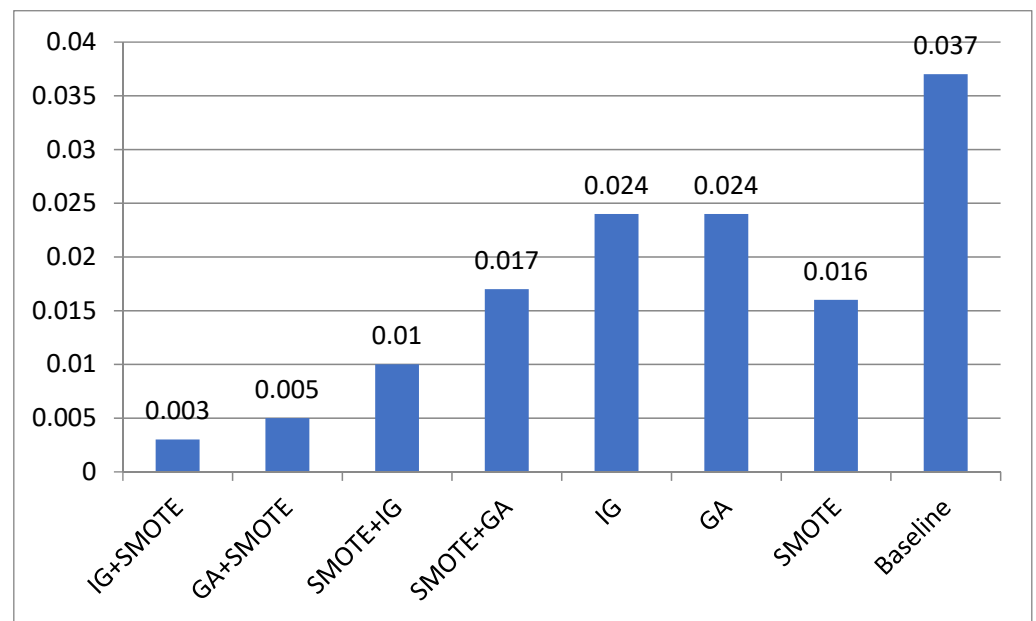


Figure 2. The type I errors of different approaches over the KDD Cup 2008 breast cancer dataset.

As we can see, the combinations of feature selection and over-sampling can allow the SVM to provide higher AUC rates and related lower type I errors than the ones with feature selection and over-sampling alone at the baseline. More specifically, the combination order of performing feature selection first and over-sampling second outperforms the opposite combination order. In particular, IG+SMOTE is the best combined approach, which causes the SVM to provide an AUC rate of 0.788 and a type I error rate of 0.003, which significantly outperforms the others ($p < 0.05$). On the other hand, for the feature reduction result, using IG and GA produce the selection of 94 and 14, respectively.

4.2. The Breast Cancer Wisconsin Dataset

Figures 3 and 4 show the AUC rates and the type I errors of different approaches, respectively. Different from the previous results, the approach that performed the best for the AUC was SMOTE (i.e., 0.962), whereas the second one was the baseline (i.e., 0.960). On the other hand, the approach that performed the best for the type I error is SMOTE+IG (i.e., 0.032), whereas the second-best ones are the baseline and SMOTE (i.e., 0.037). The other approaches producing similar AUC results were IG (i.e., 0.959), IG+SMOTE (i.e., 0.957) and SMOTE+IG (i.e., 0.955), whereas IG+SMOTE and IG produced similar type I errors, which were 0.038 and 0.044. These approaches do not have a significant level of performance difference. In particular, for the feature reduction result, using IG and GA produce 8 and 1 selected features, respectively.

The experimental results based on two different breast cancer datasets indicate that when the collected dataset is highly class imbalanced and contains a certain number of features, it is better to consider the combination of feature selection and over-sampling. Particularly, performing feature selection first and over-sampling second is likely to cause the classifier to provide higher accuracy than performing over-sampling first and feature selection second.

On the other hand, if the imbalance ratio of the collected dataset is not very high and it does not contain a large number of features, there is no need to consider the combination of feature selection and over-sampling. On the contrary, performing over-sampling to re-balance the dataset is enough to allow the classifier to provide relatively good performance.

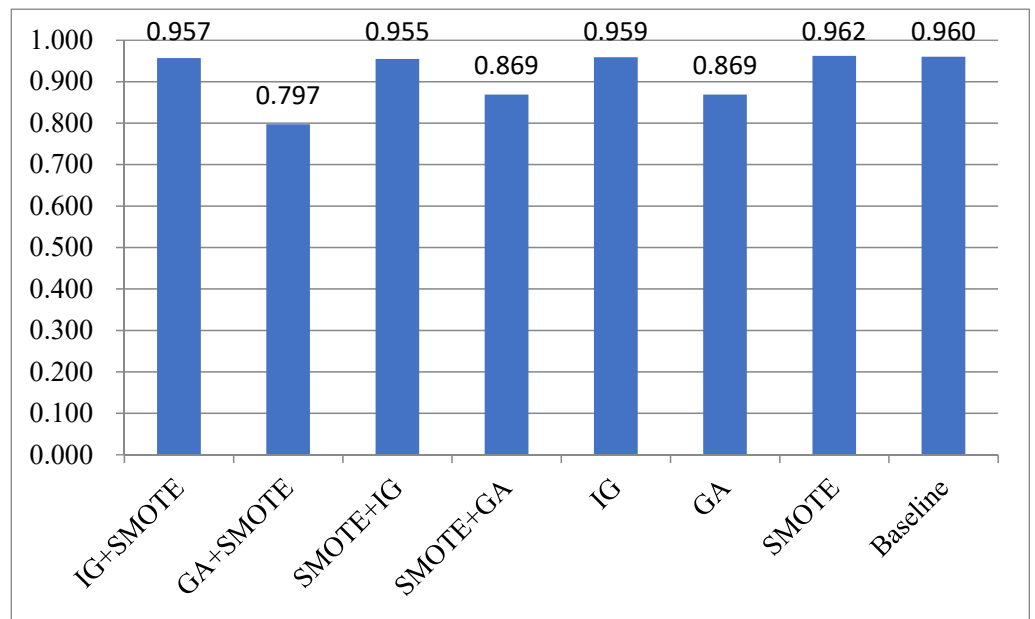


Figure 3. AUC rates of different approaches over the Breast Cancer Wisconsin Dataset.

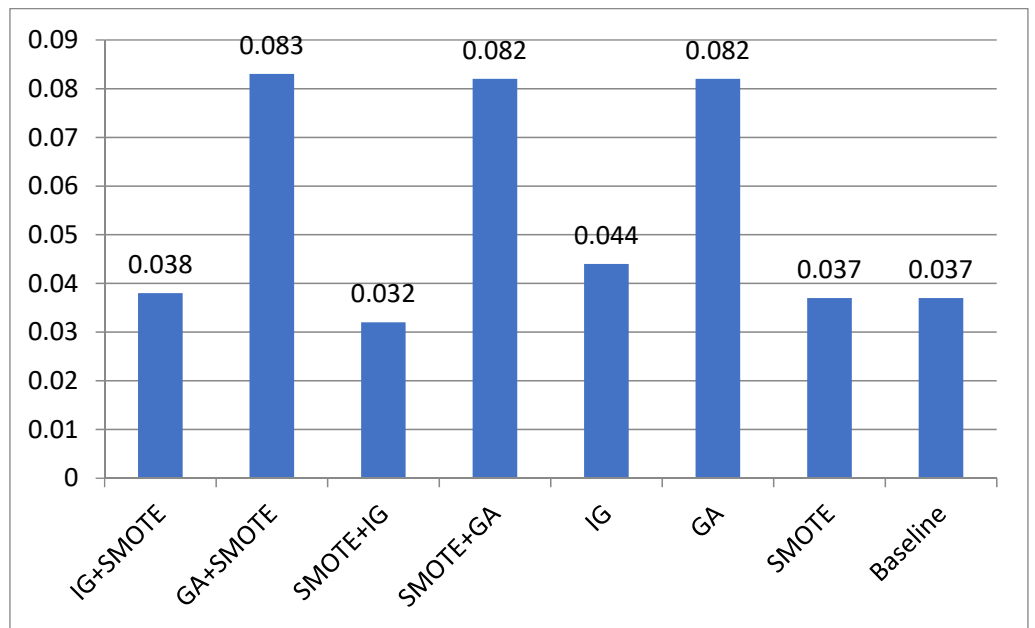


Figure 4. The type I errors of different approaches over the Breast Cancer Wisconsin Dataset.

5. Conclusions

Feature selection aims at selecting representative features from a given training set, whereas over-sampling is for re-balancing the class imbalanced training set. In this paper, the order of combining feature selection and over-sampling for breast cancer prediction are compared in terms of classification accuracy. In order to assess the performances of different combination approaches, the information gain (IG) and the genetic algorithm (GA) as the filter and wrapper-based feature selection methods and the synthetic minority over-sampling technique (SMOTE) were employed for creation of the combinations. Moreover, two breast cancer datasets with significantly different class imbalance ratios and number of features were used for the experiments.

Regarding the experimental results, for the highly imbalanced dataset containing a large number of features, performing both feature selection and over-sampling can cause

the SVM classifier provide higher AUC rates than performing feature selection and over-sampling alone as well as at the baseline. In particular, it is recommended to execute feature selection first and over-sampling second. On the contrary, for the dataset with the low imbalance ratio and small number of features, performing over-sampling alone is the better choice.

Author Contributions: Conceptualization, C.-H.C. and M.-W.H.; methodology, C.-H.C. and C.-F.T.; software, W.-C.L.; validation, C.-F.T., W.-C.L., and M.-W.H.; formal analysis, C.-H.C., C.-F.T., and M.-W.H.; resources, C.-H.C.; data curation, C.-F.T.; writing—original draft preparation, C.-H.C., C.-F.T., and M.-W.H.; writing—review and editing, M.-W.H. and C.-H.C.; supervision, C.-H.C.; project administration, M.-W.H.; funding acquisition, W.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of Taiwan, grant MOST 109–2410-H-182–012 and Chang Gung Memorial Hospital, Linkou, grant BMRPH13.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in <https://www.kdd.org/kdd-cup/view/kdd-cup-2008> (accessed on 15 February 2021) and <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29> (accessed on 15 February 2021).

Acknowledgments: The work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 109–2410-H-182–012 and in part by Chang Gung Memorial Hospital, Linkou under Grant BMRPH13.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aydinler, A.; Igcı, A.; Soran, A. *Breast Cancer: A Guide to Clinical Practice*; Springer: Berlin, Germany, 2019.
2. Zhang, Y.-D.; Satapathy, S.C.; Guttery, D.S.; Gorriz, J.M.; Wang, S.-H. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Inf. Process. Manag.* **2021**, *58*, 102439. [[CrossRef](#)]
3. Zhang, Y.-D.; Pan, C.; Chen, X.; Wang, F. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *J. Comput. Sci.* **2018**, *27*, 57–68. [[CrossRef](#)]
4. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [[CrossRef](#)]
5. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [[CrossRef](#)] [[PubMed](#)]
6. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev.* **2012**, *42*, 463–484. [[CrossRef](#)]
7. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [[CrossRef](#)]
8. Sasikala, S.; Balamurugan, S.A.A.; Geetha, S. A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer. *Procedia Comput. Sci.* **2015**, *50*, 16–23. [[CrossRef](#)]
9. Alickovic, E.; Subasi, A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Comput. Appl.* **2017**, *28*, 753–763. [[CrossRef](#)]
10. Jiang, Q.; Jin, M. Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression. *Front. Genet.* **2021**, *12*, 629946. [[CrossRef](#)] [[PubMed](#)]
11. Raj, S.; Singh, S.; Kumar, A.; Sarkar, S.; Pradhan, C. Feature selection and random forest classification for breast cancer disease. In *Data Analytics in Bioinformatics*; Wiley: Hoboken, NJ, USA, 2021; pp. 191–210.
12. Zhang, J.; Chen, L.; Tian, J.-X.; Abid, F.; Yang, W.; Tang, X.-F. Breast Cancer Diagnosis Using Cluster-based Undersampling and Boosted C5.0 Algorithm. *Int. J. Control. Autom. Syst.* **2021**, *19*, 1998–2008. [[CrossRef](#)]
13. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **2019**, *90*, 103089. [[CrossRef](#)] [[PubMed](#)]
14. Cai, T.; He, H.; Zhang, W. Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method. *Appl. Comput. Math.* **2018**, *7*, 146. [[CrossRef](#)]
15. Rani, K.U.; Ramadevi, G.N.; Lavanya, D. Performance of synthetic minority oversampling technique on imbalanced breast cancer data. In Proceedings of the 3rd International Conference on Computing for Sustainable Global Development, New Delhi, India, 16–18 March 2016; pp. 1623–1627.

16. Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [[CrossRef](#)]
17. Solanki, Y.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. *Electronics* **2021**, *10*, 699. [[CrossRef](#)]
18. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
19. Dash, M.; Liu, H. Feature selection for classification. *Intell. Data Anal.* **1997**, *1*, 131–156. [[CrossRef](#)]
20. Pintas, J.T.; Fernandes, L.A.F.; Garcia, A.C.B. Feature selection methods for text classification: A systematic literature review. *Artif. Intell. Rev.* **2021**, 1–52. [[CrossRef](#)]
21. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; De Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. *IEEE Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1106–1119. [[CrossRef](#)]
22. Davis, J.J.; Clark, A.J. Data preprocessing for anomaly based network intrusion detection: A review. *Comput. Secur.* **2011**, *30*, 353–375. [[CrossRef](#)]
23. Liang, D.; Tsai, C.-F.; Wu, H.-T. The effect of feature selection on financial distress prediction. *Knowl.-Based Syst.* **2015**, *73*, 289–297. [[CrossRef](#)]
24. Balogun, A.O.; Basri, S.; Abdulkadir, S.J.; Hashim, A.S. Performance analysis of feature selection methods in software defect prediction: A search method approach. *Appl. Sci.* **2019**, *9*, 2764. [[CrossRef](#)]
25. Tsai, C.-F.; Eberle, W.; Chu, C.-Y. Genetic algorithms in feature and instance selection. *Knowl.-Based Syst.* **2013**, *39*, 240–247. [[CrossRef](#)]
26. Huang, M.-W.; Chen, C.-W.; Lin, W.-C.; Ke, S.-W.; Tsai, C.-F. SVM and SVM ensembles in breast cancer prediction. *PLoS ONE* **2017**, *12*, e0161501.
27. Kamel, S.R.; Yaghoubzadeh, R.; Kheirabadi, M. Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. *J. Big Data* **2019**, *6*, 1–15. [[CrossRef](#)]
28. Vidić, I.; Egnell, L.; Jerome, N.P.; Teruel, J.R.; Sjøbakk, T.E.; Østlie, A.; Fjøsne, H.E.; Bathen, T.F.; Goa, P.E. Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study. *J. Magn. Reson. Imaging* **2017**, *47*, 1205–1216. [[CrossRef](#)]
29. Wang, H.; Zheng, B.; Yoon, S.W.; Ko, H.S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* **2018**, *267*, 687–699. [[CrossRef](#)]