Multicategory Survival Outcomes Classification via Overlapping Group Screening Process Based on Multinomial Logistic Regression Model With Application to TCGA Transcriptomic Data

Cancer Informatics
Volume 23: 1–17
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351241286710
Sage

Jie-Huei Wang[1], Po-Lin Hou[1] and Yi-Hau Chen[2]

[1]Department of Mathematics, National Chung Cheng University, Chiayi City, Taiwan. [2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

ABSTRACT

OBJECTIVES: Under the classification of multicategory survival outcomes of cancer patients, it is crucial to identify biomarkers that affect specific outcome categories. The classification of multicategory survival outcomes from transcriptomic data has been thoroughly investigated in computational biology. Nevertheless, several challenges must be addressed, including the ultra-high-dimensional feature space, feature contamination, and data imbalance, all of which contribute to the instability of the diagnostic model. Furthermore, although most methods achieve accurate predicted performance for binary classification with high-dimensional transcriptomic data, their extension to multi-class classification is not straightforward.

METHODS: We employ the One-versus-One strategy to transform multi-class classification into multiple binary classification, and utilize the overlapping group screening procedure with binary logistic regression to include pathway information for identifying important genes and gene-gene interactions for multicategory survival outcomes.

RESULTS: A series of simulation studies are conducted to compare the classification accuracy of our proposed approach with some existing machine learning methods. In practical data applications, we utilize the random oversampling procedure to tackle class imbalance issues. We then apply the proposed method to analyze transcriptomic data from various cancers in The Cancer Genome Atlas, such as kidney renal papillary cell carcinoma, lung adenocarcinoma, and head and neck squamous cell carcinoma. Our aim is to establish an accurate microarray-based multicategory cancer diagnosis model. The numerical results illustrate that the new proposal effectively enhances cancer diagnosis compared to approaches that neglect pathway information.

CONCLUSIONS: We showcase the effectiveness of the proposed method in terms of class prediction accuracy through evaluations on simulated synthetic datasets as well as real dataset applications. We also identified the cancer-related gene-gene interaction biomarkers and reported the corresponding network structure. According to the identified major genes and gene-gene interactions, we can predict for each patient the probabilities that he/she belongs to each of the survival outcome classes.

KEYWORDS: Multicategory classification, multinomial logistic regression, overlapping group screening, precision medicine, TCGA

RECEIVED: June 25, 2024. **ACCEPTED:** September 5, 2024.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the grant NSTC 112-2118-M-194-003-MY2 from the National Science and Technology Council of Republic of China (Taiwan). The funding body did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Jie-Huei Wang, Department of Mathematics, National Chung Cheng University, No. 168, Sec. 1, University Road, Minhsiung, Chiayi City 621301, Taiwan. Email: jhwang@ccu.edu.tw

Introduction

Precision medicine represents a cutting-edge approach to disease prevention and treatment by considering individual differences in genetics, environment, and lifestyle. In this context, cancer classification using microarray gene expression profiling has garnered significant attention.[1,2] Many statistical and machine-learning techniques have been applied to binary cancer classification using gene expression data.[3,4] These methods include multiple logistic regression models (MLRs), support vector machines (SVMs), *K*-nearest neighbors (KNNs), linear discriminant analysis (LDA), and random forests (RFs), among others. While machine learning methods are popular, they often face a major drawback: they can be difficult to interpret and may not provide direct estimates of outcome probabilities. In contrast, statistical MLRs offer both explanatory power and probabilistic estimates.

Over the past decade, multicategory classification problems have become a significant focus for biologists and computer science researchers.[5-7] In precision medicine, multicategory classification of cancer patients' survival outcomes is particularly crucial.[8,9] By employing multicategory classification, it is possible to achieve more accurate diagnoses of cancer survival outcomes, which, in turn, enables the development of more tailored and effective treatment options for patients.

The complexity of cancer development is well acknowledged, frequently involving multiple biomarkers that interact synergistically, such as gene-environment (G-E) or gene-gene (G-G) interactions.[10] Therefore, in addition to primary genetic

(G) or environmental (E) factors, interacting biomarkers can significantly influence cancer diagnosis. Incorporating these crucial interacting biomarkers into cancer classification models may improve their predictive accuracy.[11,12] However, identifying gene-gene (G-G) interactions is challenging due to the ultrahigh dimensionality of transcriptomic data. One approach to address this challenge is to utilize biological network information to help pinpoint genuine G-G interactions.[13] Additionally, another challenge is that gene expression data are often contaminated by outliers.

In high-dimensional statistical learning, regularized regression methods are commonly recommended.[14] However, a notable drawback of this approach is that the model size might exceed the sample size, potentially leading to suboptimal statistical power.[15] To address this issue, it is widely recognized that preliminary feature screening can significantly improve the effectiveness of model selection using regularization methods. Wang and Chen,[16] along with Wang et al,[17] developed overlapping group screening (OGS) methods aimed at identifying active gene-gene (G-G) and gene-environment (G-E) interactions. These methods incorporate gene pathway information and use the identified features to build a survival time prediction model. The OGS approach has also been applied to clinical cancer versus normal outcome classification using a binary logistic regression model.[4] The OGS methods are especially effective in tackling the challenge of a feature set that greatly exceeds the sample size, particularly when the feature groupings (pathways) overlap.

In this study, inspired by the methodology described by Feng et al,[9] we integrate survival indices with clinical characteristics to classify three distinct survival outcome categories: dead with no tumor, dead with tumor, and alive. This classification is applied to cancer transcriptomic data from The Cancer Genome Atlas (TCGA), specifically for kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), and head and neck squamous cell carcinoma (HNSCC). For example, the TCGA KIRP transcriptomic dataset includes 275 subjects, with 235 (85.5%) alive, 12 (4.4%) dead with no tumor, and 28 (10.0%) dead with tumor. The dataset is inherently "imbalanced" in nature.

A dataset is termed "imbalanced" when certain classes have significantly fewer subjects compared to others. This imbalance can distort classification accuracy, resulting in poor performance for minority classes despite high accuracy in majority classes. Consequently, classification models trained on imbalanced data are at a higher risk of severe overfitting and bias issues.[18] The problem of class imbalance is more pronounced in multi-class classification than in binary classification.

Several strategies have been proposed to address the class imbalance problem and develop accurate prediction models.[19,20] In this study, we focus on resampling methods, which fall into 2 main categories: over-sampling and under-sampling. Over-sampling methods involve creating synthetic samples to increase the number of instances in the minority classes. The advantage of over-sampling is that it preserves all original information, but it can lead to overfitting since it involves duplicating data from minority classes. This issue can be mitigated through techniques like cross-validation. Under-sampling methods, on the other hand, reduce the number of instances in the majority classes to balance the dataset. This approach is more effective when there is a large amount of data and the minority class is not excessively small. However, it risks losing valuable data from the majority classes, which is a significant drawback. In this work, we employ over-sampling methods to address the class imbalance issue in TCGA transcriptomic data applications.

In this study, we apply the Overlapping Group Screening (OGS) method to TCGA cancer data with multiple (>2) survival outcomes. Specifically, the OGS technique is used to identify critical transcriptomic features and gene-gene (G-G) interactions associated with these survival categories. Based on these insights, we construct microarray-based cancer diagnosis models. Unlike the traditional binary logistic regression model used by Wang and Chen,[4] we employ multinomial logistic regression to handle the multicategory outcomes. Additionally, we address the challenges posed by the ultra-high dimensionality of the gene expression data, contamination by outliers, and imbalanced outcome classes. We conduct a series of simulations to compare the performance of several machine learning methods (SVMs, LDA, RFs, and KNNs) and a penalized multinomial logistic regression model with a grouped lasso penalty against our proposed method in accurately distinguishing clinical survival samples. We apply the OGS method to TCGA cancer transcriptomic data to identify significant gene-gene (G-G) interactions associated with clinical survival categories. Based on these identified interactions, we then construct microarray-based cancer diagnosis models.

Methods

Data structure and the multiple pathways

Given a multiple $(K+1)$-class data with n subjects, where $K \geq 2$, assume that each subject i in the data belongs to a certain outcome class, so that subject i's outcome $y_i \in \{0,1,\ldots,K\}$. Also, suppose that data on subject i's p genes $\boldsymbol{x_i} = \left(x_{i1},\cdots,x_{ip}\right)'$ are available by some genotype encoding method, and the corresponding two-way gene-gene interactions are denoted by $\boldsymbol{w_i} = \left(x_{i1}x_{i2},\cdots,x_{i1}x_{ip},x_{i2}x_{i3},\cdots,x_{ip-1}x_{ip}\right)'$. Indeed, it's common for the number of genes to exceed the sample size in transcriptomic studies, and the genes are assigned to several gene pathways that may overlap with one another; that is, a given gene may belong to multiple pathways. Pathway information delineates the inherent hierarchy of genes, with overlapping pathways frequently present in gene expression data. Our objective is to uncover key genes and their interactions correlated with various clinical survival outcomes in cancer patients, utilizing pathway information to enrich our investigation. Pathway information is accessible through the Human Molecular Signature Database (MSigDB),[21] downloadable from the website http://www.broadinstitute.org/gsea/msigdb. TCGA gene expression data can be obtained from either the R package "UCSCXenaTools."[22]

Random oversampling example (ROSE) for imbalanced data

A popular over-sampling scheme for dealing with imbalanced data is random oversampling example (ROSE) proposed by Menardi and Torelli.[23] The ROSE procedure is a sampling method based on data synthesis, which addresses the problem of class imbalance by generating artificial data from a few minority classes. It recommends using model estimation and evaluation to create a more balanced data, where model evaluation is performed using a smoothed bootstrap re-sampling to validate the chosen estimation technique. The ROSE procedure can be implemented by the R package "ROSE," and can be naturally applied to the class imbalance problem in multiclass classification.

Evaluation criteria for multicategory classification

Some multicategory classification evaluation criteria are used. Let the recall REC_j and precision PRE_j for each class $j\,(j = 0,\ldots,K)$ be given as

$$REC_j = \frac{TP_j}{TP_j + FN_j} \text{ and } PRE_j = \frac{TP_j}{TP_j + FP_j} \text{ respectively,}$$

where

$$TP_j = \sum_{i=1}^{n}\left(y_i = j, \hat{y}_i = j\right); FP_j = \sum_{i=1}^{n}\left(y_i \neq j, \hat{y}_i = j\right);$$
$$FN_j = \sum_{i=1}^{n}\left(y_i = j, \hat{y}_i \neq j\right).$$

Liu et al[24] proposed the overall accuracy (OA) measure, defined as

$$OA = \frac{\sum_{j=0}^{K} TP_j}{\sum_{j=0}^{K} TP_j + \sum_{j=0}^{K} FN_j},$$

which measures the fraction of correctly classified samples over all samples, and is dominated by the performance in the majority classes. In addition, consider

$$REC = \frac{1}{(K+1)}\sum_{j=0}^{K} REC_j$$

and

$$PRE = \frac{1}{(K+1)}\sum_{j=0}^{K} PRE_j,$$

and the macro-F-measure is defined as

$$F = 2 \times \frac{PRE \times REC}{PRE + REC}.$$

In principle, higher values of REC, PRE, and F reflect better performance of the method, and, in contrast to OA, these metrics reflect more performance in the minority classes.

The overlapping group screening (OGS) approach for binary classification

Here we briefly review the OGS for binary classification in Wang and Chen.[4] This procedure involves a two-stage group screening process aimed at identifying main and interaction effects for binary classification. Considering that gene pathways may overlap with each other, that is, different pathways may share common genes, the latent effect approach proposed by Jacob et al[25] is used to consider overlapping group information. We give a simple example in the appendix to illustrate the latent effects approach, which expresses the characteristic effect of the genes as the sum of group-specific effects. All transcriptomic signatures need to be standardized before OGS methods can be applied. The procedure of the OGS method for binary logistic regression models is as follows.

Step 1: We utilize the overlapping group binary logistic regression model to identify important gene groups (pathways) by executing the R package "grpregOverlap."[26] At this stage, assume that P candidate pathways are identified among all S pathways.

Step 2: We follow the idea of Wang and Chen[16] to construct groups of G-G interaction pairs within a candidate pathway, between 2 distinct candidate pathways identified in Step 1, and between a pathway identified in Step 1 and an uncharacterized pathway. The Sequence Kernel Association Test (SKAT) by Wu et al[27] is then applied to the binary outcomes to derive group-specific P-values for each group of G-G interactions. The SKAT statistic under the binary logistic regression model is defined as

$$Q_{(h)} = \boldsymbol{m}'\boldsymbol{R}_{(h)}\boldsymbol{W}_{(h)}\mathrm{R}'_{(h)}\boldsymbol{m}, h = 1,\ldots,H,$$

where $H = P + C_2^P + (S - P) \times P$ is the total number of interacting pathway pairs considered in Step 2, $\boldsymbol{m}$ is the vector of residuals estimated from the null logistic models for binary outcomes without considering any predictors (ie, the models with only the intercept term); $\boldsymbol{R}_{(h)} = \left[r_{(h)ij}\right]_{n\times l}$, where n is the sample size and l is the number of G-G interaction pairs in the interacting pathway pair h, $r_{(h)ij}$ is the j^{th} G-G interaction pair of i^{th} subject in the interacting pathway pair h, and $\boldsymbol{W}_{(h)}$ is a diagonal weight matrix that contains the weights (for power improvement) of the l interaction pairs in the interacting pathway pair h. Following Wu et al,[27] we consider an unsupervised weight that is defined as

$$\sqrt{W_{(h)j,j}} = Beta\left(v_j, 1, 25\right), j = 1, \ldots l; h = 1, \ldots, H,$$

where $v_j = \dfrac{Var(r_{(h).j})}{\sum_{j'} Var(r_{(h).j'})}$, hence the square of the weight is a beta probability density function with parameters 1 and 25, evaluated at the ratio of the sample variance of the *j*-th interaction in the interacting pathway pair *h* to the sample variance of all interactions in this pair.

Under the null hypothesis, it is assumed that all gene-gene interaction pairs in candidate pathway h have no effect. The SKAT statistic for each G-G interaction pairs group follows a weighted sum of chi-square distribution. The group-specific P-value is obtained from the above chi-square distribution using the Davies and Algorithm[28] method, which can be computed by the R package "CompQuadForm."[29] A smaller *P*-value indicates greater significance, thereby granting higher priority in selection.

Step 3: We adopt the approach outlined by Wang et al[17] to randomly permute the original data, creating permuted data that adhere to the null model. Re-run Step 2 to calculate the group-specific *P*-values $\left\{q_1^*, \ldots, q_H^*\right\}$ and determine the desired threshold δ by selecting the minimum value among these *P*-values $\left\{q_1^*, \ldots, q_H^*\right\}$. For obtaining a stable threshold, it is necessary to repeat the permutation process multiple times, and the median of the resulting desired thresholds is utilized as the final cutoff point. G-G interaction pairs groups are deemed significant if their corresponding p-values fall below the cutoff point. Leveraging the selected pathways and G-G interaction pairs, we employ regularized logistic regression with Ridge, Lasso,[14] or adaptive Lasso penalty[30] to construct the final microarray-based classification model. This can be accomplished using the R package "glmnet."[31]

The extension of the OGS approach for binary classification to multicategory classification

Without loss of generality, we take class 0 as the reference and then consider the multicategory logistic model $\log \dfrac{\Pr(Y_i = j)}{\Pr(Y_i = 0)} = x_i' \beta_j + w_i' \gamma_j, j = 1, \ldots, K,$ which lead to $\Pr(Y_i = 0) = \dfrac{1}{1 + \sum_{j=1}^{K} exp(x_i'\beta_j + w'i\,\gamma_j)}$ and $\Pr\left(Y_i = j\right) = \dfrac{exp\left(x_i'\beta_j + w_i' \gamma_j\right)}{1 + \sum_{j=1}^{K} exp\left(x_i'\beta_j + w_i' \gamma_j\right)}, j = 1, \ldots, K.$ Note that β and γ are the corresponding effects of major genes and gene-gene interactions, respectively.

Binary classifiers are commonly used in machine learning to develop classification rules for multi-class problems.[32] One approach to applying binary classification algorithms to multi-class scenarios involves dividing the multi-class dataset into several binary-class datasets and fitting a binary classification model to each subset. This approach includes 2 main strategies: One-versus-Rest (OvR) and One-versus-One (OvO). In the OvR strategy, all classes except the one under consideration are combined into a single class, while in the OvO strategy, the model is trained to distinguish between 2 classes at a time, with each class being compared against the other classes.

The One-versus-One (OvO) strategy offers the advantage of faster training speed because each classifier is trained on data from only 2 classes, making the training process quicker than when training on all classes simultaneously. Additionally, OvO can achieve higher classification accuracy as each classifier focuses specifically on distinguishing between 2 classes. On the other hand, the advantage of the OvR strategy lies in the reduced number of classifiers required only $(K+1)$ classifiers need to be trained, with each classifier tasked to distinguish one class from all others. The implementation process is relatively straightforward, as it involves constructing $(K+1)$ binary classifiers and comparing the output of each classifier. However, its disadvantage is that each classifier may face imbalanced datasets during training, which could affect its performance. Moreover, OvR may not achieve the same high accuracy as OvO, as each classifier needs to distinguish all other classes, which can be challenging.

Following Li et al,[33] we adopt the OvO binary classifier to split a multi-class dataset into multiple binary-class datasets, and fit *K* individual binary logistic models to model the probability ratio of class *j* to class 0, *j* = 1,. . .,*K*. Accordingly, the OGS approach can be extended naturally to multinomial logistic regression models for multicategory survival outcomes in cancer diagnosis via the OvO strategy of Li et al.[33] Specifically, we divide the whole dataset into *K* datasets $\{C_1, \ldots, C_K\}$, where C_l is the dataset embracing samples from the classes 0 and l. Based on the C_l dataset, apply the OGS approach with binary logistic regression proposed by Wang and Chen[4] to obtain the corresponding estimate $\widehat{\beta}_l$ of β_l and $\widehat{\gamma}_l$ of γ_l. Repeat the above procedure from class 1 through class *K*, we can then collect $\{\widehat{\beta}_1, \ldots, \widehat{\beta}_K, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_K\}$ to get the predicted probability of occurrence of the *l-th* category given as

$$\widehat{p}_l = \frac{exp(x'\widehat{\beta_l} + w'\widehat{\gamma_l})}{1 + \sum_j^K exp(x'\widehat{\beta}_j + w'\widehat{\gamma_j})}, l = 1, \ldots, K$$

and $\widehat{p}_0 = 1 - \sum_{l=1}^{K} \widehat{p}_1$. The final classification is determined by the category with the highest predicted probability. The OvO strategy offers a key advantage in computational efficiency. This is because not all data are used simultaneously for model training, allowing for the use of parallel computing to accelerate the process.

The alternative classification methods

The "SIS_GROUP_LASSO" method utilizes a two-stage selection procedure,[34] where the top $n/(2\cdot\log(n))$ main predictors

Table 1. The gene group structure for the varying gene group-size data.

GROUP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Gene Size	3	3	3	5	6	6	9	9	9	15	15	15	24	24	24	36	36
Overlapping	1 1 0 2 2 2 3 3 0 5 5 0 8 8 0 12																

are selected in the first step by univariate multinomial logistic regressions with the marginal Akaike information criterion (AIC), and in the second step, we examine the interactions corresponding to the main effects selected in the first step. Then, the penalized multinomial logistic regression model with a grouped-lasso penalty for all the $K+1$ coefficients (corresponding to $K+1$ classes) for each selected biomarker is employed to build the classification. Since a grouped-lasso penalty is imposed on each biomarker, the effects of a biomarker over the outcome classes will all be zero or nonzero. The approach can be executed using the R package "glmnet."

In the machine learning (ML) framework, we first utilize unsupervised learning feature selection to pick the top $n/(2.\log(n))$ predictors with the largest absolute variation for subsequent ML analysis. Furthermore, the SVM method employs a radial basis kernel with a tuning gamma hyperparameter set to 1, 10^{-1}, 10^{-2}, 10^{-3} 10^{-4}, or 0. This can be implemented using the tune() function of the R package "e1071" to conduct a cross-validation process on a selection of models, aiming to derive the optimal prediction model. The KNN method utilizes a rectangular kernel and can be executed using the kknn() function from the R package "kknn." Additionally, we conduct a cross-validation process to identify the optimal k-nearest neighbor values within the KNN algorithm, aiming to derive the most accurate prediction model. Within the RF method framework, 2 hyperparameters, ntree and mtry, require tuning. We explored ntree values ranging from 1 to 500 and mtry values from 1 to 10. Subsequently, we conducted a cross-validation process on a selection of models to identify the optimal predictive model. This process can be implemented using the R package "randomForest." In summary, we have included Table A.1 of Appendix, which provides a detailed description of the hyperparameter settings for the ML approaches utilized in this paper.

Results

Simulation studies: Synthetic dataset with complex gene structure

We are currently conducting a numerical analysis to showcase our proposed OGS method with multinomial logistic regression. Additionally, we aim to evaluate the predictive efficacy of our method in comparison to several established machine learning approaches. Synthetic data consisting of 500 samples are utilized as the training set, with each subject's responses generated from a 3-class multinomial distribution,

$$p_j = \frac{exp\left(x_i'\beta_j + w_i'\gamma_j\right)}{1+\sum_{k=1}^{2} exp\left(x_i'\beta_k + w_i'\gamma_k\right)}, j=1,2 \text{ and } p_0 = 1-\sum_{j=1}^{2} p_j$$

with the covariates x are distributed uniformly between (−3,3) and w denotes the two-way interaction covariates. To assess the prediction accuracy of various methods, we generate a test dataset comprising 300 samples, drawn from the same distribution as the training data but independent of it. Within the framework of 3-class multinomial logistic regression, we predict the subject's label using the following equation,

$$\hat{Y} = l \; if \; max\left\{\hat{p}_0, \hat{p}_l, \hat{p}_2\right\} = \hat{p}_l$$

where $\hat{p}_k$, $k=012$, are obtained by the method in Section 2.5.

The simulation considers the gene group size (the number of genes per group) and the overlapping structure (the number of genes shared by two overlapping groups), as outlined in Table 1, where we can see, for instance, groups 10 and 11 each consist of 15 genes, totaling 25 unique genes between them, with 5 genes shared. Overall, this study encompasses 193 genes and 243 potential group-specific gene effects. Figure 1 illustrates the associated gene network structure. We further hypothesize that different biomarker effects are present in different outcome classes. In class 1, we hypothesize the efficacy of gene groups 9 and 11, with genes in each group exhibiting consistent effects of −1.5 and 1.5, respectively. Moreover, within group 9, effective G-G interactions (G37-G39, G38-G40) demonstrate effects of (1.5, 1.5), while between groups 9 and 11, effective G-G interactions (G41-G58, G42-G59) display effects of (1.5, 1.5). In class 2, we hypothesize the efficacy of gene groups 13 and 15, with genes in each group exhibiting consistent effects of 1.5 and 1.5, respectively. Moreover, within group 13, effective G-G interactions (G78-G80, G79-G81) demonstrate effects of (1.5, 1.5), while between groups 13 and 15, effective G-G interactions (G82-G118, G83-G119) display effects of (1.5, 1.5). There are 18 721 major genes and G-G interaction pairs in this simulation study, and the average proportions of outcome classes 1, 2, and 0 are 35%, 39%, and 26%, respectively.

We conducted the described simulation setup 500 times to gather numerical results. The results presented in Table 2 indicate that the OGS method employing Ridge, Lasso, and Adaptive Lasso penalties consistently outperforms other methods, including common ML techniques, in multi-class prediction.

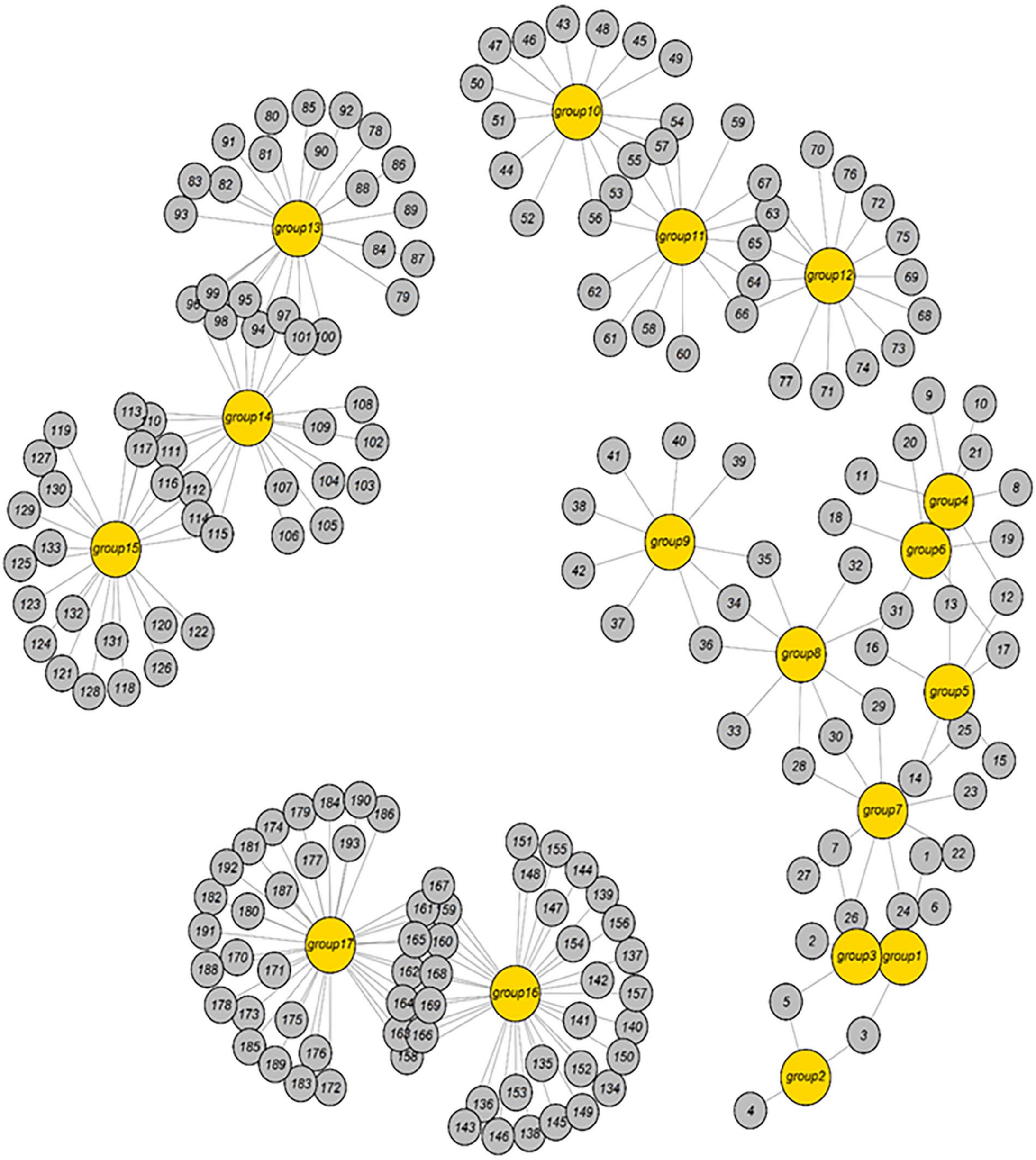

Figure 1. The gene network structure for the varying gene group-size data.

Table 2. Averages (standard deviations) of testing prediction performance over 500 simulated replicates for various multi-class classification methods under the gene structure with different gene group sizes.

METHODS	OA	PRE	REC	*F*
OGS_Ridge	0.7067 (0.0433)	0.6987 (0.0449)	0.6983 (0.0441)	0.6985 (0.0444)
OGS_Lasso	0.7028 (0.0401)	0.6951 (0.0416)	0.6947 (0.0410)	0.6949 (0.0412)
OGS_ALasso	0.6923 (0.0390)	0.6853 (0.0401)	0.6849 (0.0401)	0.6850 (0.0400)
SIS_GROUP_LASSO	0.4271 (0.0399)	0.4054 (0.0465)	0.4012 (0.0396)	0.4031 (0.0420)
SVM	0.4763 (0.0346)	0.4636 (0.0361)	0.4565 (0.0334)	0.4600 (0.0344)
LDA	0.4786 (0.0351)	0.4669 (0.0352)	0.4645 (0.0343)	0.4657 (0.0346)
RF	0.4720 (0.0345)	0.4667 (0.0507)	0.4349 (0.0309)	0.4496 (0.0375)
KNN	0.4662 (0.0309)	0.4567 (0.0356)	0.4414 (0.0285)	0.4487 (0.0308)

Table 3. The gene group structure for the equal gene group-size data.

GROUP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Gene Size	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Overlapping	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3																	

We also explore an alternate gene network structure, comprising 18 groups, each containing 10 genes. Details regarding group sizes and overlapping structure are provided in Table 3. This example encompasses 129 genes and 180 potential group-specific gene effects. Figure 2 illustrates the associated gene network structure. In class 1, we hypothesize the efficacy of gene groups 1 and 4, with genes in each group exhibiting consistent effects of −1.5 and 1.5, respectively. Moreover, within group 1, effective G-G interactions (G1-G3, G2-G4) demonstrate effects of (1.5, 1.5), while between groups 1 and 4, effective G-G interactions (G5-G25, G6-G26) display effects of (1.5, 1.5). In class 2, we hypothesize the efficacy of gene groups 13 and 18, with genes in each group exhibiting consistent effects of 1.5 and 1.5, respectively. Moreover, within group 18, effective G-G interactions (G123-G125, G124-G126) demonstrate effects of (1.5, 1.5), while between groups 13 to 18, effective G-G interactions (G88-G127, G89-G128) display effects of (1.5, 1.5). There are 8385 major genes and G-G interaction pairs in this simulation study, and the average proportions of outcome classes 1, 2, and 0 are 38%, 36%, and 26%, respectively.

From the results presented in Tables 2 and 4, it's apparent that the OGS method with Ridge, Lasso, and Adaptive Lasso penalties consistently outperforms other methods, including traditional ML approaches, in terms of classification performance. Additionally, both Tables 2 and 4 showcase the standard deviations of accuracy metrics across different methods, indicating that the OGS methods exhibit slightly higher variability in accuracy compared to alternative approaches.

Real data application: Kaplan-Meier survival curves

We first display the 3 Kaplan-Meier survival curves for the 3 cancer types (KIRP, LUAD, and HNSCC) across the 3 groups (alive, dead with no tumor, and dead with tumor). We then perform a log-rank test to assess whether there are significant differences between the survival curves of these 3 groups. From Figure 3, it can be observed that there are significant differences in the survival curves among the 3 groups in the survival data of KIRP and HNSCC. However, in the survival data of LUAD, there are no significant differences in the survival curves between the "dead with no tumor" and "dead with tumor" groups.

Real data application: TCGA KIRP data

Our own TCGA KIRP data consist of 275 subjects, of whom 235 (85.5%) alive, 12 (4.4%) dead with no tumor, and 28 (10.0%) dead with tumor. The data is extremely imbalanced in terms of the outcome class distribution. Given that the pool of cancer-related genes is likely finite, it makes sense to streamline the gene set before constructing the classification model. We employ unsupervised learning for feature selection, identifying the top 1000 genes with the most significant absolute variation for subsequent analysis.

For the proposed OGS approach, out of the initial 1000 genes selected through unsupervised learning, 697 genes are linked to 398 pathways based on prior pathway information from the GO Cellular Component (GO-CC) database. The remaining 303 genes, not mapped to any pathway in the GO-CC database, are either excluded or grouped together in the OGS method. These alternative approaches result in a total of 243 253 and 500 500 main and G-G interaction effects, respectively.

We randomly split the entire dataset into 10 sets of 165:110 for 60% training and 40% testing, respectively, to evaluate the performance of all considered methods. The ROSE resampling is performed on the training data to address the class imbalance issue. Table 5 summarizes the average 10-fold classification results after removing 303 ungrouped genes from the analysis. We also consider another pathway database, Kyoto Encyclopedia of Genes and Genomes (KEGG),[35-37] and the corresponding analysis results are shown in Table A.2 of Appendix. From both sets of results, we see that the OGS method has better classification performance compared to the other methods in terms of *REC*, *PRE*, and *F* performance metrics. The ML methods SVM and KNN have superior performance in terms of the metric *OA* but inferior *REC*, *PRE*, and *F* metrics compared to the OGS, owing to that the ML methods perform well in the majority outcome classes (alive and dead with tumor outcomes), but perform poorly in the minority class (dead with no tumor outcome).

Next, based on the GO-CC database, we apply the OGS approach with the adaptive lasso penalty to the entire TCGA KIRP data, and examine the selected features in the dead with tumor outcome category. The method selects 95 G-G interaction biomarkers, and the corresponding network is shown in Figure 4. Some selected biomarkers have been shown to have biological meaningful in published literature. For example, Wang et al[38]

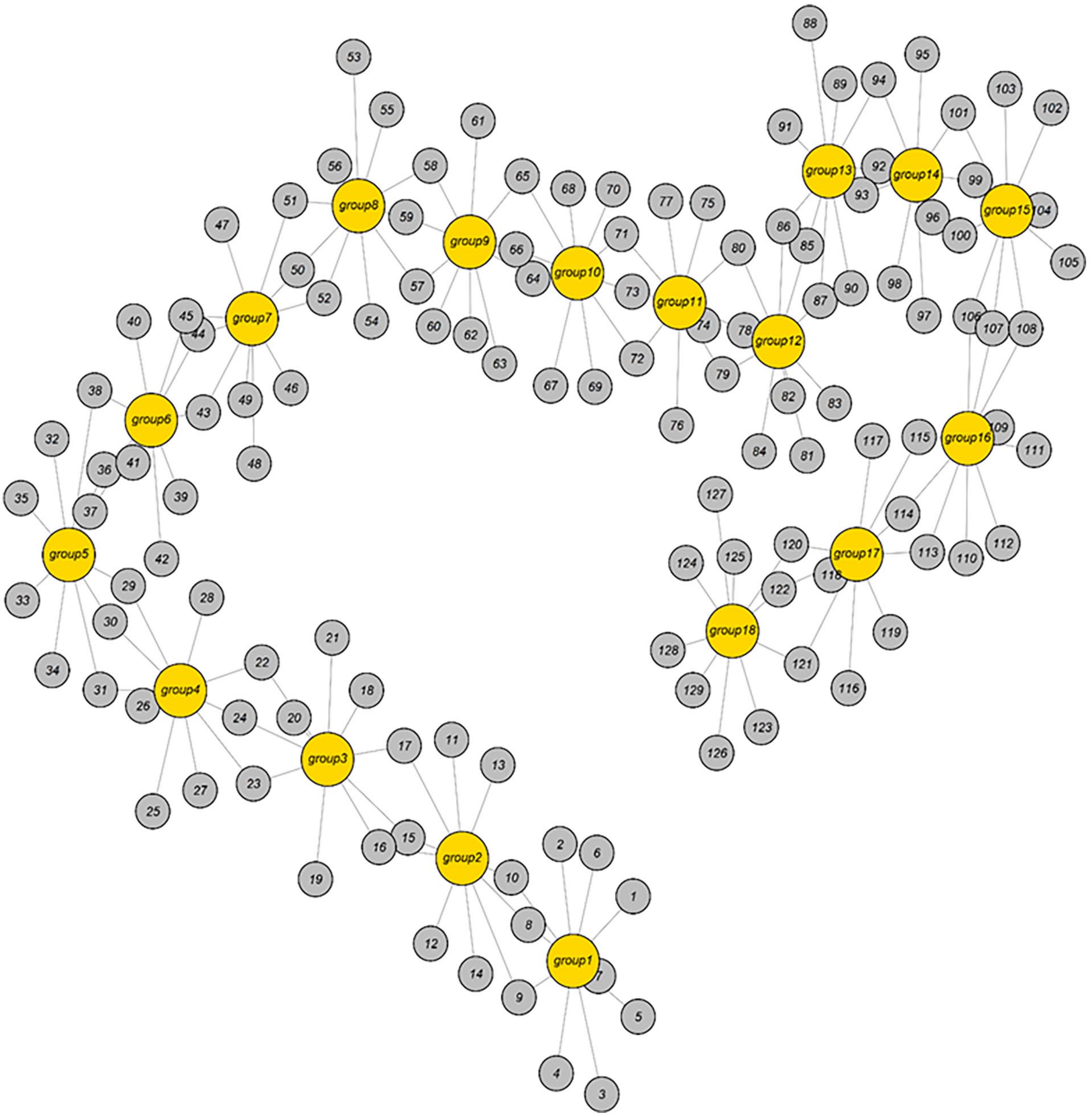

Figure 2. The gene network structure for the equal gene group-size data.

Table 4. Averages (standard deviations) of testing prediction performance over 500 simulated replicates for various multi-class classification methods under the gene structure with equal gene group sizes.

METHODS	OA	PRE	REC	*F*
OGS_Ridge	0.6777 (0.0387)	0.6707 (0.0405)	0.6694 (0.0400)	0.6701 (0.0402)
OGS_Lasso	0.7065 (0.0379)	0.7000 (0.0394)	0.6988 (0.0391)	0.6994 (0.0392)
OGS_ALasso	0.7124 (0.0425)	0.7067 (0.0442)	0.7051 (0.0439)	0.7059 (0.0440)
SIS_GROUP_LASSO	0.5410 (0.0449)	0.5249 (0.0505)	0.5145 (0.0442)	0.5196 (0.0471)
SVM	0.5813 (0.0343)	0.5734 (0.0356)	0.5646 (0.0351)	0.5689 (0.0350)
LDA	0.5894 (0.0341)	0.5805 (0.0350)	0.5773 (0.0349)	0.5789 (0.0348)
RF	0.5561 (0.0339)	0.5666 (0.0428)	0.5235 (0.0340)	0.5439 (0.0358)
KNN	0.5326 (0.0297)	0.5302 (0.0357)	0.5107 (0.0304)	0.5201 (0.0312)

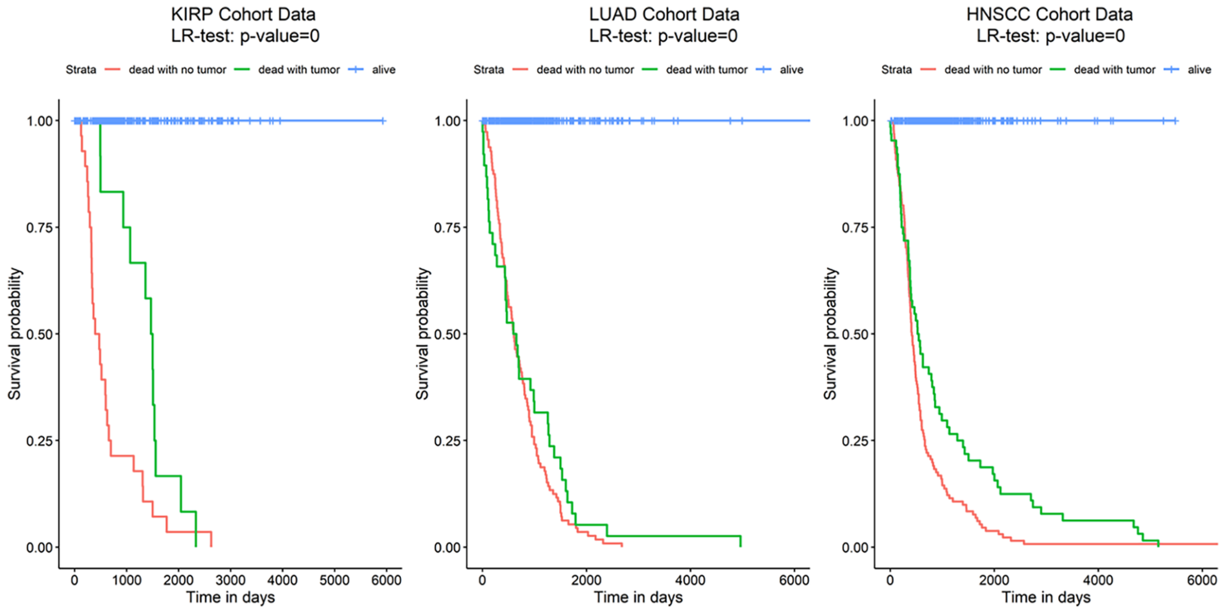

Figure 3. Kaplan-Meier survival outcomes for the three cancer types (KIRP, LUAD, and HNSCC) across the three groups (alive, dead with no tumor, and dead with tumor).

Table 5. Averages (standard deviations) of testing prediction performance of different methods with GO_CC gene sets databases in the TCGA KIRP data over 10 random splits of 165:110 training/test sets.

METHODS	OA	PRE	REC	*F*
OGS_Ridge	0.6136 (0.0372)	0.4242 (0.0384)	0.5237 (0.0467)	0.4680 (0.0381)
OGS_Lasso	0.6282 (0.0292)	0.4288 (0.0337)	0.5192 (0.0482)	0.4689 (0.0344)
OGS_ALasso	0.6173 (0.0375)	0.4272 (0.0337)	0.5170 (0.0428)	0.4669 (0.0309)
SIS_GROUP_LASSO	0.4064 (0.0284)	0.4094 (0.0304)	0.5606 (0.0841)	0.4721 (0.0497)
SVM	0.6782 (0.0567)	0.3809 (0.0534)	0.4060 (0.0851)	0.3908 (0.0649)
LDA	0.6445 (0.0723)	0.3790 (.0.281)	0.4046 (0.0528)	0.3909 (0.0386)
RF	0.6482 (0.0828)	0.3900 (0.0508)	0.4211 (0.0556)	0.4047 (0.0516)
KNN	0.8464 (0.0347)	0.4090 (0.1048)	0.3809 (0.0624)	0.3928 (0.0794)

showed that the "*HOXDs*" gene is lowly expressed in KIRP, and the upregulation of "*HOXDs*" is associated with improved overall survival of cancer patients. These findings suggested that "*HOXDs*" may be an indicator biomarker for pan-cancer prognosis and immunotherapy. Jia et al[39] demonstrated the expression and function of "*CAMK2B*" in vitro and in vivo, and provided evidence that this protein promotes reregulation of the stromal tumor microenvironment and inhibits KIRP proliferation.

Real data application: TCGA LUAD data

The TCGA LUAD data consist of 454 subjects, of whom 304 (67.0%) alive, 38 (8.4%) dead with no tumor, and 112 (24.7%) dead with tumor. There exists class imbalance in this dataset. We choose the top 1000 genes with the highest absolute variation for subsequent analysis.

For the proposed OGS approach, out of the initial 1000 genes selected through unsupervised learning, 640 genes are linked to 402 pathways based on prior pathway information from the GO-CC database. The remaining 360 genes, not mapped to any pathway in the GO-CC database, are either excluded or grouped together in the OGS method. These alternative approaches result in a total of 205 120 and 500 500 main and G-G interaction effects, respectively. We randomly split the entire dataset into 10 sets of 272:182 for 60% training and 40% testing, respectively, to evaluate the performance of all considered methods. The ROSE resampling is performed on the training data to alleviate the class imbalance.

Table 6 summarizes the average 10-fold classification results after removing 360 ungrouped genes from the analysis. We also consider KEGG pathway database, and the corresponding analysis results are shown in Table A.3 of Appendix. From both sets

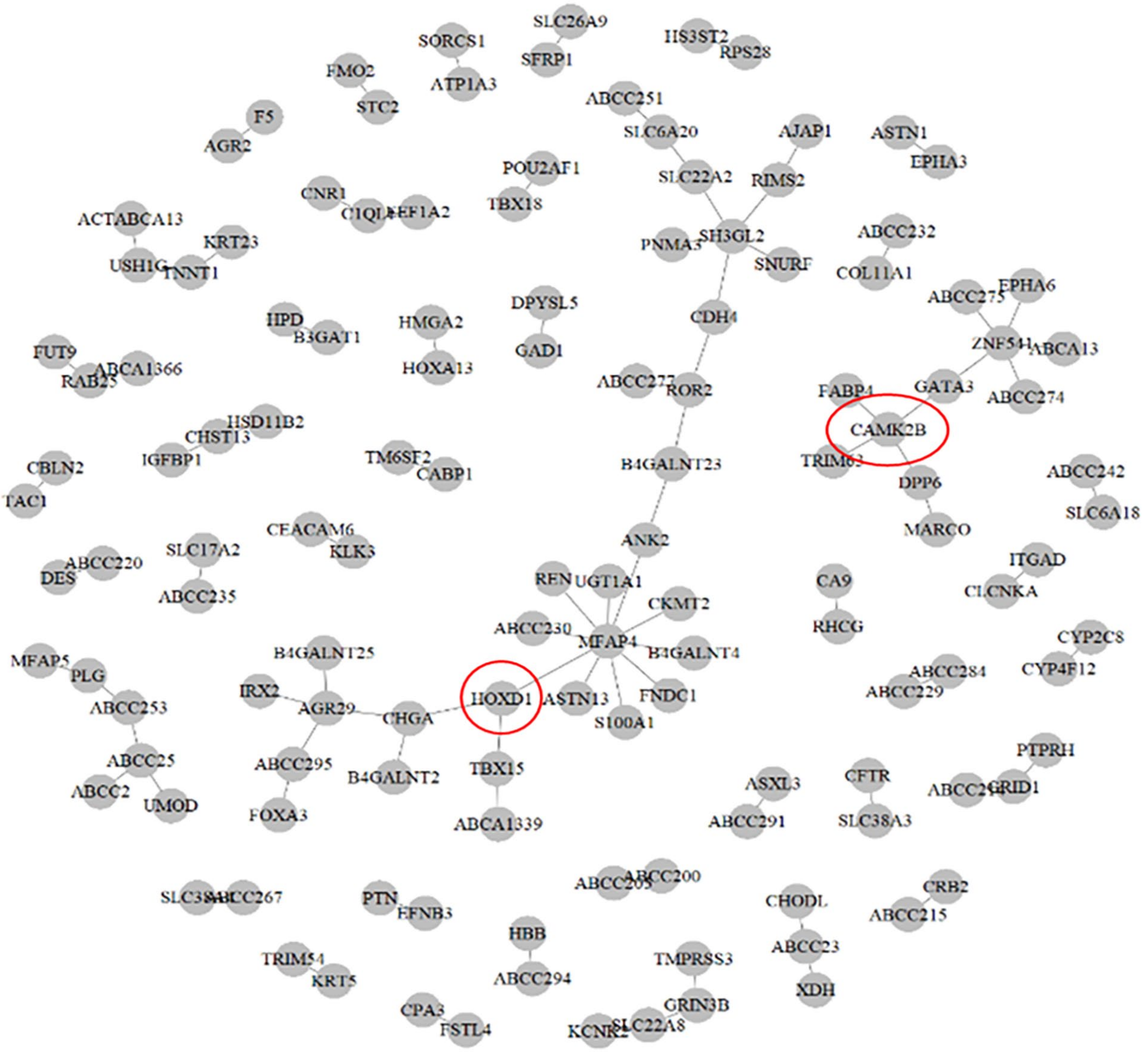

Figure 4. The network of the selected G-G interactions by the OGS approach with the adaptive lasso penalty in the TCGA KIRP gene expression data with dead with tumor outcome.

Table 6. Averages (standard deviations) of testing prediction performance of different methods with GO_CC gene sets databases in the TCGA LUAD data over 10 random splits of 272:182 training/test sets.

METHODS	OA	PRE	REC	*F*
OGS_Ridge	0.5879 (0.0272)	0.4161 (0.0357)	0.4054 (0.0391)	0.4105 (0.0364)
OGS_Lasso	0.5621 (0.0328)	0.4006 (0.0383)	0.3974 (0.0627)	0.3986 (0.0500)
OGS_ALasso	0.5154 (0.0335)	0.3824 (0.0315)	0.3894 (0.0513)	0.3855 (0.0399)
SIS_GROUP_LASSO	0.3874 (0.0545)	0.3525 (0.0313)	0.3476 (0.0480)	0.3497 (0.0396)
SVM	0.6368 (0.0163)	0.3124 (0.1021)	0.3272 (0.0125)	0.3138 (0.0515)
LDA	0.4753 (0.0292)	0.3300 (0.0392)	0.3349 (0.0462)	0.3322 (0.0414)
RF	0.2593 (0.0708)	0.3752 (0.0203)	0.3552 (0.0346)	0.3646 (0.0272)
KNN	0.6582 (0.0243)	0.3518 (0.0291)	0.3439 (0.0094)	0.3473 (0.0160)

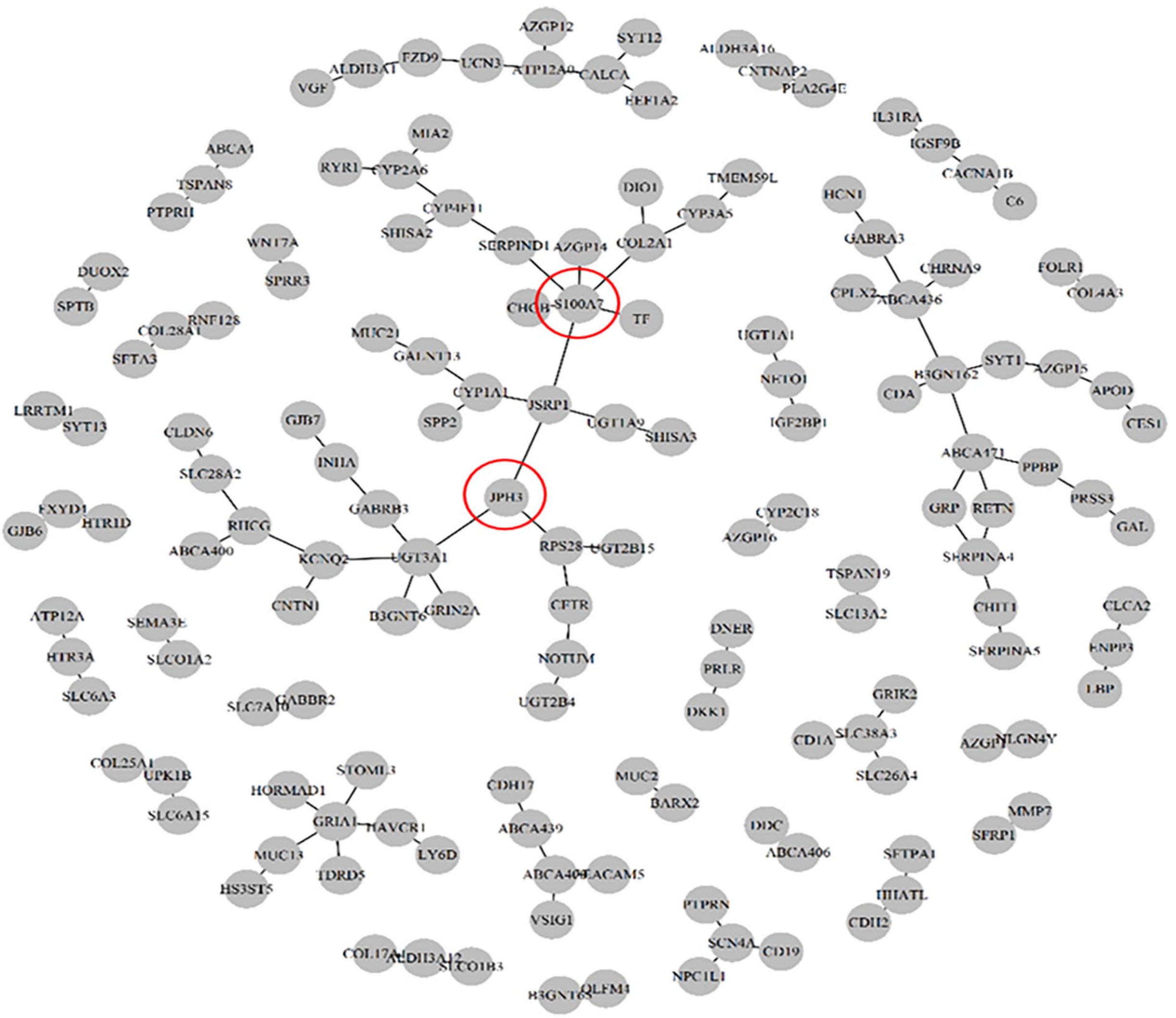

Figure 5. The network of the selected G-G interactions by the OGS approach with the adaptive lasso penalty in the TCGA LUAD gene expression data.

of results, it is apparent that the OGS method consistently exhibits superior classification performance in terms of *REC*, *PRE*, and *F* metrics compared to other methods. The SVM and KNN have superior performance in terms of the metric *OA* but inferior *REC*, *PRE*, and *F* metrics compared to the OGS, owing to that the ML methods perform well in the majority outcome classes (alive and dead with tumor outcomes), but perform poorly in the minority class (dead with no tumor outcome).

Based on the GO-CC database, the OGS approach with the adaptive lasso penalty selects 121 G-G interaction biomarkers, and the corresponding network is shown in Figure 5. Some selected biomarkers have been shown to have biological meaningful in published literature. For example, Zhang et al[40] showed that the gene "*JPH3*" was associated with non-small cell lung cancer (NSCLC), and they found that some genes including "*JPH3*" were frequently silenced by epigenetic mechanisms in lung cancer. Also, Nasser et al[41] demonstrated that "*S100A7*" is upregulated in multiple types of malignancies, including non-small cell lung cancer, contributing to tumor growth, premetastatic niche formation, and metastasis.

Real data application: TCGA HNSCC data

The TCGA HNSCC data consist of 491 subjects, of whom 296 (60.3%) alive, 64 (13.0%) dead with no tumor, and 131 (26.7%) dead with tumor. The data is moderately imbalanced in terms of the outcome class distribution. We choose the top 1000 genes with the highest absolute variation for subsequent analysis.

For the proposed OGS approach, out of the initial 1000 genes selected through unsupervised learning, 667 genes are linked to 393 pathways based on prior pathway information from the GO-CC database. The remaining 333 genes, not mapped to any pathway in the GO-CC database, are either excluded or grouped together in the OGS method. These alternative approaches result in a total of 222 778 and 500 500 main and G-G interaction effects, respectively. We randomly split the entire dataset into 10 sets of 295:196 for 60% training and 40% testing, respectively, to evaluate the performance of all considered methods. The ROSE resampling is performed on the training data to alleviate the class imbalance.

Table 7. Averages (standard deviations) of testing prediction performance of different methods with GO_CC gene sets databases in the TCGA HNSCC data over 10 random splits of 295:196 training/test sets.

METHODS	OA	PRE	REC	*F*
OGS_Ridge	0.4924 (0.0301)	0.3678 (0.0292)	0.3669 (0.0331)	0.3674 (0.0311)
OGS_Lasso	0.4848 (0.0234)	0.3856 (0.0285)	0.3813 (0.0323)	0.3834 (0.0301)
OGS_ALasso	0.4640 (0.0309)	0.3894 (0.0301)	0.3866 (0.0352)	0.3879 (0.0323)
SIS_GROUP_LASSO	0.3843 (0.0436)	0.3792 (0.0392)	0.3721 (0.0385)	0.3756 (0.0386)
SVM	0.5756 (0.0514)	0.3397 (0.0837)	0.3523 (0.0346)	0.3426 (0.0572)
LDA	0.4538 (0.0572)	0.3636 (0.0468)	0.3621 (0.0510)	0.3628 (0.0488)
RF	0.3142 (0.0453)	0.3600 (0.335)	0.3457 (0.0311)	0.3524 (0.0303)
KNN	0.5761 (0.0335)	0.4193 (0.1382)	0.3592 (0.0288)	0.3797 (0.0642)

Table 7 summarizes the average 10-fold classification results after removing 333 ungrouped genes from the analysis. We also consider KEGG pathway database, and the corresponding analysis results are shown in Table A.4 of Appendix. These results reveal that, the OGS approach has slightly better performance metrics than the ML methods in terms of *REC*, *PRE*, and *F* classification metrics, which focus more on the rare class, while the ML method SVM has the best classification performance in terms of the OA metric, which focuses more on the dominant classes.

Based on the GO-CC database, the OGS approach with the adaptive lasso penalty selects 85 G-G interaction biomarkers, and the corresponding network is shown in Figure 6. Some selected biomarkers have been shown to have biological meaningful in published literature. For example, Irimie et al[42] showed that the gene "*MGST1*" was associated with HNSCC, and they found that the expression levels of several genes, including the "*MGST1*" gene, were altered between smoking and nonsmoking HNSCC patients. Misawa et al[43] showed that neuropeptide genes including "*GAL*" are powerful epigenetic biomarkers in HNSCC.

We also report the biomarkers with the top- and bottom-coefficients for the "dead with tumor" outcome, identified using the OGS approach with adaptive lasso penalty to the 3 entire TCGA transcriptomic data, based on the GO-CC and KEGG databases. The results are detailed in Table 8 and A.5 of Appendix. Positive coefficients indicate that higher biomarker expression increases "dead with tumor" event probability, while negative coefficients indicate it decreases event probability.

Additionally, we plot the receiver operating characteristic (ROC) curves for the "dead with tumor" category of these 3 real datasets using various classification methods, and calculated the corresponding area under curve (AUC) values for both training and testing data. The ROC curves and AUC values for all methods are obtained by averaging the results from 10 iterations of the validation set approach. The corresponding graphs are shown in Figure 7 and A.1 of Appendix. From the ROC curves and AUC values in these 2 figures, we can conclude that our proposed method effectively avoids the overfitting problem compared to the common machine learning methods considered in the article.

The advantage of our proposed method compared to machine learning approaches is that we emphasize model inference. Specifically, we focus on understanding the relationship between important biomarkers and the response variable, not black box models. In terms of prediction, our method allows us to calculate the probability that an observation belongs to a certain class, rather than merely predicting a classification. ROC curve analysis of real data also indicates that our method can avoid overfitting, which is a common issue with machine learning methods.

Discussion

In summary, we outline the similarities and distinctions among Wang et al,[17] Wang and Chen,[4] and the current paper. Wang et al[17] employed the OGS approach with Cox's regression model to identify significant gene-environment interactions linked to clinical censoring survival outcomes. Conversely, Wang and Chen[4] utilized the OGS approach with a binary logistic regression model to discover critical gene-gene interaction biomarkers associated with the occurrence of binary cancer/normal outcomes. This article employs the OvO strategy to convert multi-class classification into multiple binary classifications. It then integrates this approach with the OGS procedure, utilizing a binary logistic regression model as outlined by Wang and Chen.[4] This combination aims to identify significant gene-gene interaction biomarkers associated with multiple survival statuses in cancer patients. In this paper, in addition to the typical challenges of ultra-high dimensionality and feature contamination, we also encounter the problem of data imbalance. Together, these factors pose significant obstacles to accurate predictive modeling.

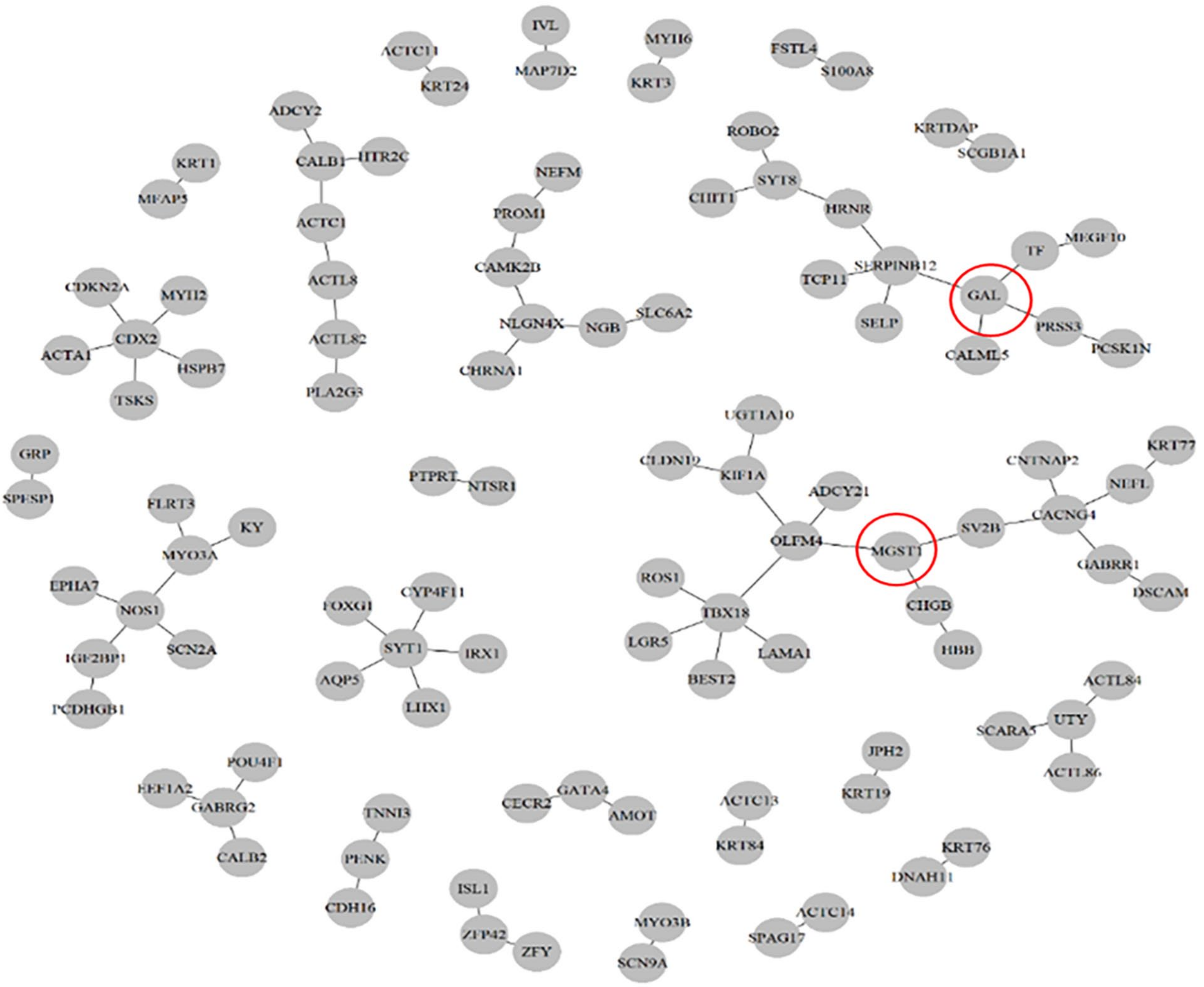

Figure 6. The network of the selected G-G interactions by the OGS approach with the adaptive lasso penalty in the TCGA HNSCC gene expression data.

Potential improvements to the OGS method

Since the real-world data we are interested in is imbalanced, there are 3 main approaches to dealing with class imbalance: resampling, cost-sensitive, and ensembling, and several extensions based on these approaches have also been developed.[19,44-46] In the real data applications, we just leverage the ROSE resampling procedure to balance the data, while a remaining interesting problem is, how to find a best way to tackle class imbalance for downstream genome-wide association study (GWAS), and how this way may work with the OGS approach.

In practical data analysis, we first select the top 1000 genes with the highest variance in gene expression. However, since variance itself is susceptible to outliers and gene data is often contaminated, it is essential to explore more suitable unsupervised feature selection methods.[47] Moreover, Fan and Lv[15] pointed out marginal feature selection may overlook key predictors due to: (1) Joint correlation not captured by marginal analysis, (2) Selection of secondary predictors highly correlated with important ones, and (3) Collinearity among predictors. They proposed an iterative method to address these issues. In addition, feature selection is the process of trying to select more informative features. Too many redundant or irrelevant features may overwhelm the important features of the classification. Feature selection can solve such problems, thereby improving prediction accuracy and reducing the computational cost of classification algorithms. Another interesting issue is that after the OGS procedure selects the most important genes and gene pairs, we can try feeding these selected biomarkers into another machine learning algorithm to see how well the predictions perform. We will study these further issues in future work.

In the OGS method, the SKAT test is key for screening gene interactions. Lee et al[48] evaluated various gene- or region-based testing methods, including burden and variance-component tests, and assessed their performance. Since different methods have unique strengths based on the biological context, future research should explore diverse testing approaches to enhance OGS effectiveness. The OGS method extracts gene network information using predefined pathways, which limits it to genes in those pathways and can result in information loss. Researching ways to relax these constraints could improve feature selection and classification prediction. Besides, we utilize

Table 8. Biomarkers with the top- and bottom-coefficients for the "dead with tumor" outcome are identified using the OGS approach with the adaptive lasso, based on GO-CCdatabase.

KIRP (95 ACTIVE)		LUAD (121 ACTIVE)		HNSCC (85 ACTIVE)	
ID	COEFFICIENT	ID	COEFFICIENT	ID	COEFFICIENT
Top 10 biomarkers					
RIMS2-SH3GL2	0.735	GABRB3-INHA	0.560	ROS1-TBX18	0.516
MFAP4-ABCC230	0.557	APOD-AZGP15	0.533	KRT1-MFAP5	0.495
CLCNKA-LILRA4	0.479	CYP2A6-RYR1	0.474	DNAH11-KRT76	0.472
CHGA-AGR29	0.359	CYP1A1-SPP2	0.395	KIF1A-OLFM4	0.452
KRT5-TRIM54	0.344	CYP4F11-SHISA2	0.309	HRNR-SYT8	0.372
CRB2-RHCG	0.315	ABCA439-ABCA490	0.297	MEGF10-TF	0.353
HBB-ABCC294	0.293	AZGP15-SYT1	0.294	CALML5-GAL	0.348
ABCC2-ABCC25	0.274	ABCA471-PPBP	0.284	GRP-SPESP1	0.314
CBLN2-TAC1	0.271	AZGP1-NLGN4Y	0.263	SELP-SERPINB12	0.307
GPRC5A-KCNH6	0.251	GRIA1-MUC13	0.255	CAMK2B-NLGN4X	0.263
Bottom 10 biomarkers					
GRID1-ABCC214	−0.312	GRIA1-TDRD5	−0.294	CDX2-HSPB7	−0.191
DES-ABCC220	−0.327	DKK1-PRLR	−0.346	ACTL82-PLA2G3	−0.210
SH3GL2-SNURF	−0.340	CHIT1-SERPINA5	−0.350	ACTC1-ACTL8	−0.231
SH3GL2-SLC22A2	−0.382	ABCA400-RHCG	−0.366	KRT3-MYH6	−0.234
GAD1-USH1G	−0.401	SLC6A15-UPK1B	−0.423	PENK-TNNI3	−0.249
CHGA-HOXD1	−0.404	MMP7-SFRP1	−0.425	SERPINB12-TCP11	−0.254
ASXL3-ABCC291	−0.419	PTPRH-TSPAN8	−0.445	CHIT1-SYT8	−0.386
KRT23-TNNT1	−0.429	ALDH3A1-FZD9	−0.496	CAMK2B-PROM1	−0.439
POU2AF1-TBX18	−0.523	CHGB-S100A7	−0.539	NEFM-PROM1	−0.454
CDH4-ABCC274	−0.633	ALDH3A1-VGF	−0.564	KRT77-NEFL	−0.520

two-way and multiplicative interactions for simplicity in interaction assessments. However, higher-order and more complex interactions are challenging and warrant further research.

Misclassification analysis discussing

In this study, based on the classification by Feng et al,[9] we divided cancer patients into 3 groups according to their survival status and clinical condition (ie, dead due to cancer, dead due to other reasons, and alive). However, this classification could be further refined, for example: (1) Survivors Close to Death: If some survivors are in a health condition very close to death, this might impact the accuracy and interpretation of the classification results. Introducing indicators of health severity could improve classification accuracy. (2) Death without Tumor Population: If some individuals may not have been diagnosed with tumors or if the cause of death records are inaccurate, this could affect the accuracy of the analysis. To mitigate this impact, using grading or other indicators to assess the actual condition of these patients could be considered. (3) Censoring Issues: In survival data, censoring is an important consideration. Future research should incorporate censoring information into the cancer patient classification to achieve more accurate classification and prediction. These considerations could help improve the precision of the classification and the reliability of the analysis, leading to a better understanding of cancer patients' survival conditions and treatment outcomes.

Multiple cancer subtypes classification

van't Veer and Bernards[49] highlighted that identifying cancer subtypes is crucial for personalized precision medicine, as treatment decisions heavily depend on understanding these subtypes. For instance, Lavagna et al[3] developed a biomarker predictor to classify small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), while Tian et al[50] used a

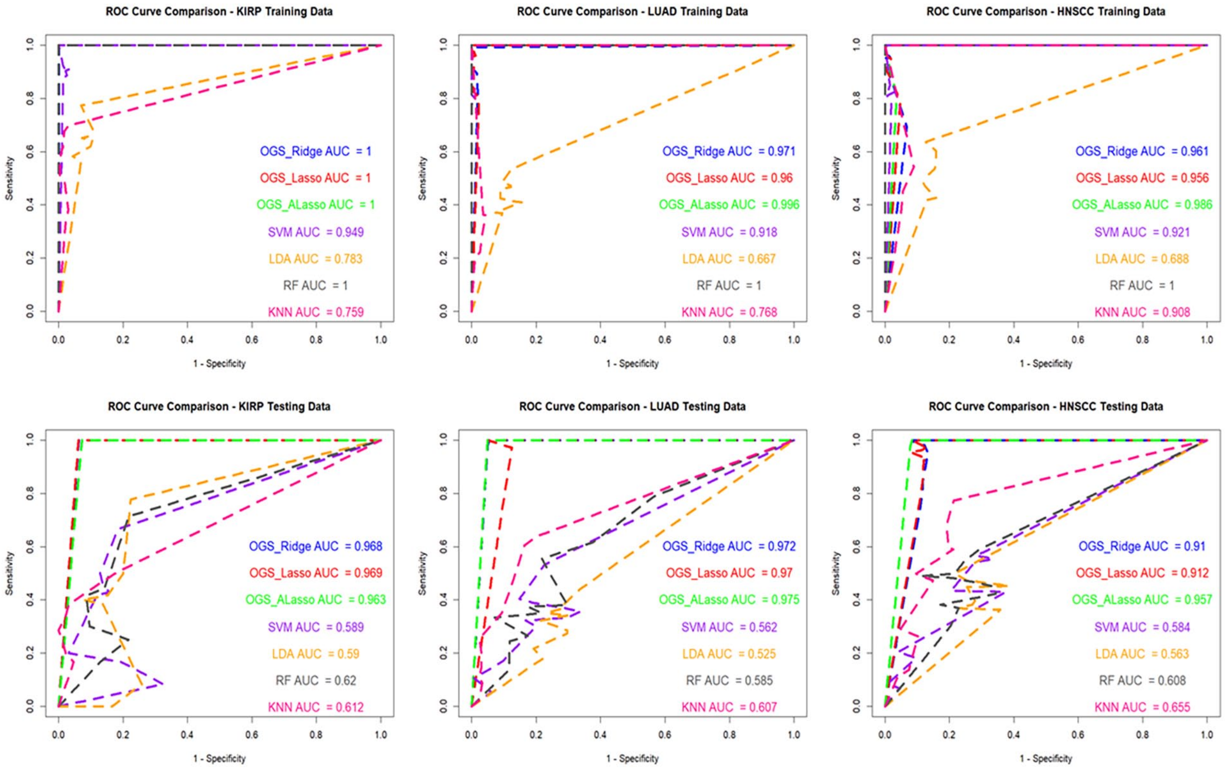

Figure 7. Average ROC curves and AUC values for "dead with tumor" across KIRP, LUAD, and HNSCC datasets, evaluated with different classification methods on training and test data, based on GO-CC database.

network-constrained sparse multinomial logit model to predict glioblastoma multiforme (GBM) subtypes. Tabibu et al[51] applied deep learning to classify renal cell carcinoma and predict survival outcomes based on pathological images. Such classification methods support early diagnosis, enabling more precise treatments and prognostic assessments. Consequently, the OGS with multinomial logistic regression model could also be used for cancer subtype identification and prediction.

State-of-the-art methods

Recent studies at the forefront of the field have demonstrated improved classification results. It is widely recognized that metaheuristic algorithms have been extensively utilized to enhance classification performance. For example, the hybridization of Particle Swarm Optimization has improved crime rate prediction.[52] Similarly, combining Cuckoo Search with Harris Hawks Optimization has boosted cancer detection rates,[53] while integrating Cuckoo Search with deep learning has enhanced cancer disease classification.[54] Additionally, Marine Predator Chaotic Search has proven effective for detecting COVID-19.[55] Therefore, we are going to conduct a series of investigations and studies to examine the performance of different state-of-the-art methods on multiclass imbalanced biological data.

Large datasets integrations

Several public human databases, including Gene Expression Omnibus (GEO), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), National Cancer Database (NCDB), and The Cancer Genome Atlas (TCGA), are valuable for assessing the reproducibility of our findings. We believe that conducting a meta-analysis could help discover and validate survival prognostic biomarkers[56] and plan to explore this in future research.

Conclusion

In this article, we employ the OvO strategy to transform multiclass classification into multiple binary classification, and utilize the OGS with binary logistic regression to include gene pathway information for identifying important major genes and gene-gene interactions for multicategory survival outcomes. Based on the identified biomarkers, we can predict for each patient the probabilities that he/she belongs to each of the outcome classes. In simulation studies, we demonstrate that the classification performance of our proposed method outperforms some commonly used ML methods and the multinomial logistic regression with the group lasso penalty. In real data applications, we employ the ROSE resampling procedure to address the class imbalance and analyze 3 sets of TCGA cancer transcriptomic data (KIRP, LUAD, and HNSCC). The numerical results demonstrate that the new proposal leads to a substantial improvement in cancer diagnosis when compared to methods that do not take pathway information into account.

Acknowledgements

The authors sincerely appreciate Ms. Zih-Han Wu for her assistance with creating figures and organizing tables during

the article revision process. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

JH conceived and designed the experiments. PL collected and organized the analysis data. JH and PL analyzed the data. JH and YH wrote the first draft of the manuscript. JH and YH made critical revisions and approved final version. All authors agreed with manuscript results and conclusions. All authors jointly developed the structure and arguments for the paper. All authors reviewed and approved of the final manuscript.

Availability of Data and Materials

The R codes for both simulation studies and real data applications can be accessed on the figshare website: https://doi.org/10.6084/m9.figshare.23849679.v2. Additionally, the transcriptomic data for TCGA KIRP, LUAD, and HNSCC, along with the clinical multicategory survival outcomes examined in this study, are available on figshare: https://doi.org/10.6084/m9.figshare.23849370.v3. The TCGA data analyzed in this study were originally obtained from the TCGA Hub repository: https://tcga.xenahubs.net, with the primary source being the TCGA Website: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

Ethics Approval and Consent to Participate

The study described in this manuscript did not include human or animal participants. All data used in this research were sourced from publicly available and freely accessible repositories. Therefore, ethical approval was unnecessary for this study.

Consent for Publication

Not applicable.

ORCID iD

Jie-Huei Wang https://orcid.org/0000-0003-1596-8471

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Zhu Y, Shen X, Pan W. Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*. 2009;10:1-11.
2. Rauschert S, Raubenheimer K, Melton PE, Huang RC. Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. *J Clin Epigenet*. 2020;12(1):51. doi:10.1186/s13148-020-00842-4
3. Lavanya C, Pooja S, Kashyap AH, et al. Novel biomarker prediction for lung cancer using random forest classifiers. *Cancer Inform*. 2023;22:1-15.
4. Wang JH, Chen YH. Overlapping group screening for binary cancer classification with TCGA high-dimensional genomic data. *J Bioinform Comput Biol*. 2023;21:2350013.
5. Piao Y, Piao M, Ryu KH. Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. *Comput Biol Med*. 2017;80:39-44.
6. Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform*. 2013;14:13-26.
7. Li J, Wang Y, Song X, Xiao H. Adaptive multinomial regression with overlapping groups for multi-class classification of lung cancer. *Comput Biol Med*. 2018;100:1-9.
8. Deng F, Shen L, Wang H, Zhang L. Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models. *Am J Cancer Res*. 2020;10:4624-4639.
9. Feng CH, Disis ML, Cheng C, Zhang L. Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models. *Lab Invest*. 2022;102:236-244.
10. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10:392-404.
11. Li J, Dong W, Meng D. Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information. *IEEE/ACM Trans Comput Biol Bioinform*. 2018;15:2028-2038.
12. Wang JH, Chen YH. Interaction screening by Kendall's partial correlation for ultrahigh-dimensional data with survival trait. *Bioinformatics*. 2020;36:2763-2769.
13. Wang JH, Chen YH. Network-adjusted Kendall's tau measure for feature screening with application to high-dimensional survival genomic data. *Bioinformatics*. 2021;37:2150-2156.
14. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol*. 1996;58:267-288.
15. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodol*. 2008;70:849-911.
16. Wang JH, Chen YH. Overlapping group screening for detection of gene-gene interactions: application to gene expression profiles with survival trait. *BMC Bioinform*atics. 2018;19:335.
17. Wang JH, Wang KH, Chen YH. Overlapping group screening for detection of gene-environment interactions with application to TCGA high-dimensional survival genomic data. *BMC Bioinformatics*. 2022;23:202.
18. Taha AY, Tiun S, Abd Rahman AH, Sabah A. Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *ITB J Inf Commun Technol*. 2021;20:423-456.
19. Selamat NA, Abdullah A, Mat Diah N. Association features of smote and rose for drug addiction relapse risk. *J King Saud Univ - Comput Inf Sci*. 2022;34: 7710-7719.
20. Abdoh SF, Abo Rizka M, Maghraby FA. Cervical cancer diagnosis using random forest classifier with smote and feature reduction techniques. *IEEE Access*. 2018;6:59475-59485.
21. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. 2005;102:15545-15550.
22. Wang S, Liu X. The ucscxenatools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. *J Open Source Softw*. 2019;4:1627.
23. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Disc*ov. 2014;28:92-122.
24. Liu Z, Tang D, Cai Y, Wang R, Chen F. A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. *Neurocomputing*. 2017;266:641-650.
25. Jacob L, Obozinski G, Vert J. Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp.433–440. Montreal, QC: ACM.
26. Zeng Y, Breheny P. Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Inform*. 2016;15:179-187.
27. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89:82-93.
28. Davies RB, Algorithm AS. Algorithm AS 155: the distribution of a linear combination of χ 2 random variables. *J R Stat Soc Ser C Appl Stat*. 1980;29:323-333.
29. Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput Stat Data Anal*. 2010;54:858-862.
30. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418-1429.
31. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39:1-13.
32. Aly M. Survey on multiclass classification methods. *Technical Report, Caltech;* 2005.
33. Li J, Chen Z, Wang Z, Chang YCI. Active learning in multiple-class classification problems via individualized binary models. *Comput Stat Data Anal*. 2020;145:1-17. doi:10.1016/j.csda.2020.106911
34. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*. 2009;25:714-721.
35. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27-30.

36. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci*. 2019;28:1947-1951.
37. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 2023;51:D587-D592.
38. Wang L, Wang X, Sun H, Wang W, Cao L. A pan-cancer analysis of the role of HOXD1, HOXD3, and HOXD4 and validation in renal cell carcinoma. *Aging*. 2023;15:10746-10766.
39. Jia Q, Liao X, Zhang Y, et al. Anti-tumor role of CAMK2B in remodeling the stromal microenvironment and inhibiting proliferation in papillary renal cell carcinoma. *Front Oncol*. 2022;12:1-12.
40. Zhang K, Wang J, Yang L, et al. Targeting histone methyltransferase g9a inhibits growth and Wnt signaling pathway by epigenetically regulating HP1α and APC2 gene expression in non-small cell lung cancer. *Mol Cancer*. 2018;17:153.
41. Nasser MW, Wani NA, Ahirwar DK, et al. RAGE mediates S100A7-induced breast cancer growth and metastasis by modulating the tumor microenvironment. *Cancer Res*. 2015;75:974-985.
42. Irimie AI, Braicu C, Cojocneanu R, et al. Differential effect of smoking on gene expression in head and neck cancer patients. *Int J Environ Res Public Health*. 2018;15:1558.
43. Misawa K, Mima M, Imai A, et al. The neuropeptide genes SST, TAC1, HCRT, NPY, and GAL are powerful epigenetic biomarkers in head and neck cancer: a site-specific analysis. *Clin Epigenetics*. 2018;10:52.
44. Ozçift A. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Comput Biol Med*. 2011;41:265-271.
45. Ali S, Majid A, Javed SG, Sattar M. Can-CSC-GBE: developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data. *Comput Biol Med*. 2016;73:38-46.
46. Wang JH, Liu CY, Min YR, Wu ZH, Hou PL. Cancer diagnosis by gene-environment interactions via combination of SMOTE-Tomek and overlapped group screening approaches with application to imbalanced TCGA clinical and genomic data. *Mathematics*. 2024;12:2209.
47. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. *Artif Intell Rev*. 2020;53:907-948.
48. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95:5-23.
49. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*. 2008;452:564-570.
50. Tian X, Wang X, Chen J. Network-constrained group lasso for high-dimensional multinomial classification with application to cancer subtype prediction. *Cancer Inform*. 2014;13:25-33.
51. Tabibu S, Vinod PK, Jawahar CV. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Sci Rep*. 2019;9:10509.
52. Rahman RU, Singh K, Tomar DS, et al. Building resilient digital forensic frameworks for NoSQL database: harnessing the blockchain and quantum technology. In: Kumar A, Ahuja NJ, Kaushik K, eds. *Sustainable Security Practices Using Blockchain, Quantum and Post-Quantum Technologies for Real Time Applications*. Springer; 2024:205-238.
53. Yaqoob A, Verma NK, Aziz RM, et al. Enhancing feature selection through metaheuristic hybrid cuckoo search and Harris hawks optimization for cancer classification. In: Kalita K, Ganesh N, Balamurugan S, eds. *Metaheuristics for Machine Learning*. John Wiley & Sons, Incorporated; 2024:95-134.
54. Joshi AA, Aziz RM. A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function. *Multimed Tools Appl*. 2024;83:71721-71752.
55. Saxena A, Chouhan SS, Aziz RM, Agarwal V. A comprehensive evaluation of Marine predator chaotic algorithm for feature selection of COVID-19. *Evol Syst*. 2024;15:1235-1248.
56. Liu X, Wang J, Chen M, et al. Combining data from TCGA and GEO databases and reverse transcription quantitative PCR validation to identify gene prognostic markers in lung cancer. *Onco Targets Ther*. 2019;12:709-720.