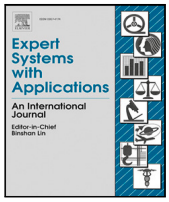
Expert Systems With Applications 226 (2023) 120179

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Biomarker discovery from high-throughput data by connected network-constrained support vector machine

Lingyu Li, Zhi-Ping Liu *

School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

ARTICLE INFO

Keywords:
Network-constrained support vector machine
Biomarker discovery
Connectivity
Feature selection
High-throughput data
Breast cancer

ABSTRACT

From a systems biology perspective, genes usually work collaboratively in the form of a network, e.g., cancer-related genes participate in an integrative dysfunctional pathway. Thus, feature gene selection considering the graph or network structure plays a crucial role in cancer biomarker discovery from high-throughput omics data. The network-based paradigm demonstrates that integrating gene expression data with gene networks can improve classification performances and generate more interpretable feature subsets. In this paper, we propose an embedded connected network-constrained support vector machine (CNet-SVM) method to keep the selected features in an inherent graph structure in discovering biomarker genes. Firstly, we mathematically formulate the CNet-SVM model as a convex optimization problem constrained by network connectivity inequalities and theoretically investigate the behaviors of all tuning parameters to provide search guidance on the regularization path. Secondly, to check if the genes selected by CNet-SVM could be studied as network-structured biomarkers, we conduct experiments on several simulation datasets and real-world breast cancer (BRCA) datasets to validate its classification and prediction capabilities. The results show that CNet-SVM not only maintains the sparsity and smoothness, but also considers the connectivity constraints between genes when selecting features on a prior gene–gene interaction network from omics data. Especially, CNet-SVM identifies 32 BRCA biomarker genes, which form into a connected network component and can be potentially used for BRCA diagnosis. Furthermore, the comparisons with eight feature selection-empowered SVM methods demonstrate that the easily interpretable networked feature genes discovered by CNet-SVM are more closely related to BRCA dysfunctions. Finally, we validate that the identified biomarkers achieve high prediction accuracy on external independent cohorts. All results proved that the proposed CNet-SVM method is effective in selecting connected-network-structured features and can be an alternative improvement to the current SVM models for biomarker identification from high-throughput data. The data and code are available at https://github.com/zpliulab/CNet-SVM.

1. Introduction

Biomarker discovery is essential for the diagnosis and prognosis of breast cancer (BRCA) (Coleto-Alcudia & Vega-Rodríguez, 2022). The increasingly developed high-throughput RNA sequencing (RNA-seq) technology has revolutionized the way people understand the genetics of BRCA (Kim et al., 2021). Several regularized sparse statistical models have been developed for the biomarker discovery of BRCA using bulk RNA-seq data (Li et al., 2020; Sarkar, Saha, Sarkar, & Maulik, 2021). Essentially, biomarker discovery in medicine is equivalent to feature selection in machine learning (Al-Obeidat, Tubaishat, Shah, Halim, et al., 2020; Li, Ching, & Liu, 2022; Li & Liu, 2020; Sarkar et al., 2021; Wei, Gu, & Zhang, 2021). In biological pathways, a chain of molecular interactions leads to new molecular product creation or alters the cellular state (Iqbal & Halim, 2020). In cells, genes tend to be functionally associated in the form of a network and contribute to the biological outcome in a synergistic manner (Kong & Yu, 2018).

Support vector machine (SVM) is a powerful machine learning method with many practical applications. In the past few years, SVM has been widely used in different bioinformatics domains, like protein phosphorylation site prediction (Lin et al., 2015), multi-location protein subcellular localization (Wan, Mak, & Kung, 2012), protein–protein interaction prediction (Cui, Fang, & Han, 2012), membrane protein function prediction (Wan, Mak, & Kung, 2016) and biomarker discovery (Coleto-Alcudia & Vega-Rodríguez, 2022). However, the standard SVM cannot automatically perform feature selection in classification, and therefore several regularized support vector machine (Reg-SVM)

* Corresponding author.
E-mail addresses: lingyuli@mail.sdu.edu.cn (L. Li), zpliu@sdu.edu.cn (Z.-P. Liu).

https://doi.org/10.1016/j.eswa.2023.120179
Received 21 March 2022; Received in revised form 8 March 2023; Accepted 15 April 2023
Available online 22 April 2023

procedures have been developed to extend its flexibility. For example, Zhu, Rosset, Tibshirani, and Hastie (2003) introduced a Reg-SVM model with L_1-norm penalty term (Lasso-SVM), which not only forces the reduction in variances of fitted coefficients, but also shrinks coefficients toward zeros for the penalty of sparseness. Then AAbraham, Kowalczyk, Zobel, and Inouye (2013) presented another Reg-SVM model with elastic net penalty (ENet-SVM) and showed that it achieved higher precision and lower false-positive rates than the other Reg-SVM methods. Besides, Zhang, Ahn, Lin, and Park (2006) developed a novel Reg-SVM model with the smoothly clipped absolute deviation (SCAD) non-convexity regularization term, i.e., SCAD-SVM, but it might be too strict in selecting features for non-sparse data (Fan & Li, 2001). Moreover, Becker, Toedt, Lichter, and Benner (2011) presented an embedded Reg-SVM model with L_2-norm and SCAD penalty, called L2SCAD-SVM, which overcomes the limitations of each penalty alone. Nonetheless, the above-existing Reg-SVM methods do not consider the interconnection structure information between genes when selecting feature genes (AAbraham et al., 2013; Jung, 2013; Wang, Shao, Zhou, Zhang, & Xiu, 2021).

In addition, the grouping structure penalty for biomarker selection ensures that the selected biomarkers are a continuous and meaningful group structure (Meier, Van De Geer, & Bühlmann, 2008) rather than a list of outliers selected by the above-mentioned Reg-SVM methods. For instance, Yuan and Lin (2006) proposed a group Lasso (GLasso) model to make feature selection on pre-defined groups of variables. Besides, Yang and Zou (2015) developed a fast algorithm for solving GLasso penalized learning problems. However, GLasso penalty only constrains sparsity between groups but no sparsity within groups (Ma, Song, & Huang, 2007). Therefore, Simon, Friedman, Hastie, and Tibshirani (2013) provided a sparse group Lasso (SGLasso) penalty model that considers the effects of both between-group and within-group variables. The SGLasso method has been proven to select more effective feature genes and performs better than the original method. Moreover, Huo et al. (2020) constructed an SGLasso-SVM method for tumor classification and investigated different feature gene selection methods and classifiers via experiments. Nevertheless, it is valuable to note that although grouping structure penalties consider the group effect variables and select more compelling features, it needs to pre-define the grouping information between features. In a word, the pre-defined gene group is mainly based on prior knowledge. For example, Sun, Fan, Lelieveldt, and van de Giessen (2015) used the brain structure segmentation as the prior knowledge of how brain voxels and the corresponding weight space should be divided into several non-overlapping groups.

Recently, the molecular network structure underlying biomarkers has attracted significant research interests (Kong et al., 2022), but there is still a gap to achieve the connected network-structured biomarkers as we aspired. For example, based on the work of Zou and Yuan (2008), Zhu, Shen, and Pan (2009) proposed a network-based SVM model with F_∞-norm to pick out pairwise gene neighbors in the sense of group variable selection. The strategy facilitates the model to extract more biological insights from gene expression data, but it still fails to select structured genes forming into a connected network component. Additionally, Becker, Werft, Toedt, Lichter, and Benner (2009) and Chen, Xuan, Riggins, Clarke, and Wang (2011) built up an integrated SVM model with network constraints to identify cancer biomarkers from gene expression data. However, the model only imposes the smoothness for coefficients, it needs a significance test to screen subnetwork genes according to P-value. Moreover, Chai, Huang, Jiang, Liang, and Xia (2016) developed another Net-SVM model combined with $L_{1/2}$-norm, which performs promising for cancer classification and gene selection on high-dimensional and small-size sample data. However, it still results in some feature genes that are isolated and independent without forming a connected network. In all these available Reg-SVM methods, the selected features do not have a connected network structure and often result in independent and/or isolated features. Thus, it leads that the biomarkers they recognized may not be functionally relevant in the application scenario of discovering disease signature genes. Therefore, if used for BRCA biomarker identification, they generate false-positive feature genes and may not correctly clarify the molecular mechanism behind the prevalence of BRCA (Jubair, Alkhateeb, Abou Tabl, Rueda, & Ngom, 2020).

Biological observations reveal that a group of genes forming in clusters, clique-like structures or neighbors in a connected network tend to function together in specific biological processes, e.g., to work together to initiate specific cancers (Tanvir, Aqila, Maharjan, Mamun, & Mondal, 2019; Zhu et al., 2009). Li and Liu (2022) demonstrated that the biomarkers discovered by network-based pipelines show more promising performances than those identified from purely data-driven approaches. Therefore, to address the limitations in the former available Reg-SVM models, we aim to propose a connected network-constrained regularized SVM model, called CNet-SVM, for feature selection and explore its ability to identify network biomarkers for BRCA by integrating transcriptome and interactome. In our model, the feature selection is implemented under the rules of connected network constraints when carrying out the classification. Thus, all biomarker genes selected by CNet-SVM model are linked on a connected network forming a component of the prior gene interaction network. The results on both simulated data and BRCA data demonstrate the improved performance over the state-of-the-art other SVM-based methods for biomarker discovery from high-throughput data. By employing the concept of connected network regularization (Li & Liu, 2022), we introduce the penalty term to the optimization functions of an SVM model. It is still a constraint optimization problem based on structural risk minimization. The major contributions of this paper are the following ones:

- The CNet penalty is firstly embedded into the standard SVM model to develop a geometry-based classification approach (CNet-SVM) to realize the discovery/identification of network biomarkers.
- CNet-SVM model considers the connectivity constraints between genes when selecting features and ensures the selected feature genes form into a connected network component. The numerical experiments on several simulation data and real BRCA data demonstrate its effectiveness and efficiency.
- The comparison study with numerous existing feature selection-empowered SVM methods shows that the CNet-SVM model unearths potential network structure and obtains better classification performances simultaneously.
- Function enrichment analysis shows that the connectivity and network topology between biomarkers provide an essential understanding of how genes perform functions cooperatively and indicate their dysfunctions in cancer mechanisms in the form of pathways.

The rest of this paper is composed of sections as follows. In Section 2, we first introduce CNet-SVM in the Reg-SVM framework and present how to select parameters and perform feature selection. Section 3 assesses the performance of our proposed CNet-SVM method. Namely, we compare it with numerous feature selection strategies for SVM, including SVM-RFE, one wrapper method based on recursive feature elimination; mRMR-SVM, one filter method based on minimum redundancy maximum relevance; four embedded Reg-SVM methods, i.e., Lasso-SVM, ENet-SVM, SCAD-SVM and L2SCAD-SVM; and two grouping structure methods: GLasso-SVM and SGLasso-SVM. We not only implement these methods on extensive simulated data to select features but also apply them to public BRCA RNA-seq data to identify potential biomarker genes. In Section 4, we discuss the related issues in this work. Finally, we conclude in Section 5 and point out our further research directions.

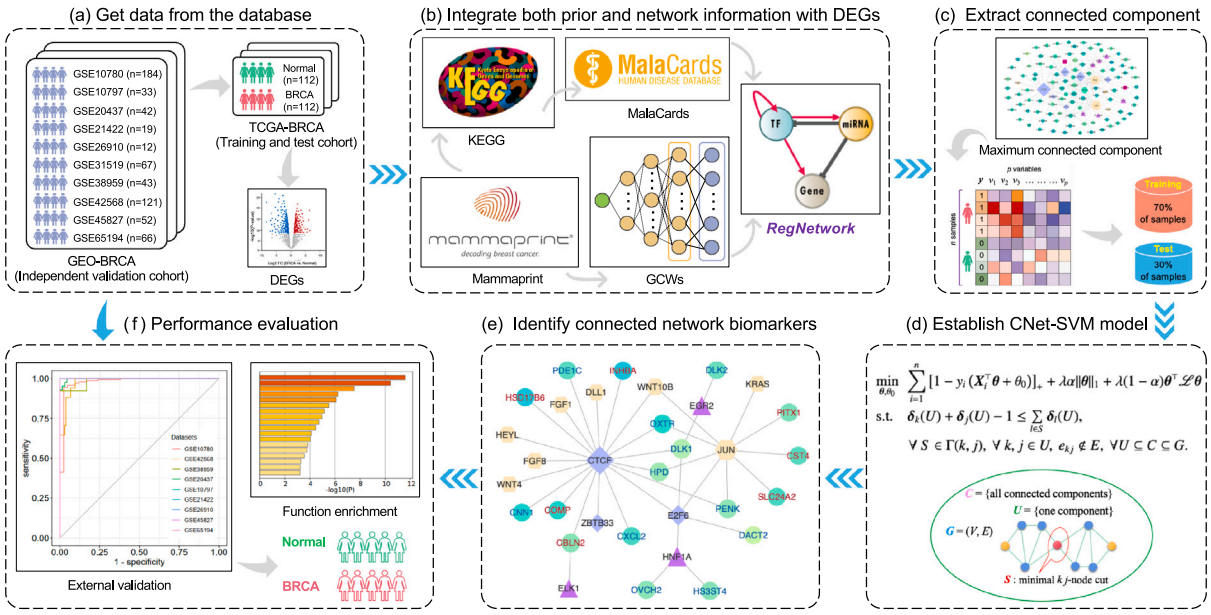


Fig. 1. The framework for discovering structural network biomarkers of BRCA from gene expression data by CNet-SVM. (a) Access the available gene expression data and select differentially expressed genes. (b) Add the prior BRCA-related knowledge from MalaCards, KEGG, MammaPrint, top-ranked genes by GCWs and some documented TFs to DEGs and link them with the gene interaction network compiled in RegNetwork. (c) Extract the maximum connected component from the integrated gene network and map the RNA-seq data of these genes in the connected component. (d) Construct the CNet-SVM model and apply it to the network-structured data to select features. (e) Identify the selected connected networking feature genes as biomarkers of BRCA. (f) Evaluate the performances regarding classification accuracy and biological function implications enriched in the selected biomarkers.

2. Materials and methods

2.1. Framework

Fig. 1 illustrates the framework for discovering structural network biomarkers of BRCA by the CNet-SVM method. Firstly, we collect gene expression data of BRCA and implement necessary data preprocessing. Secondly, we integrate prior genes and network information with differentially expressed genes (DEGs), in which the BRCA-related genes from MalaCards (Rappaport et al., 2017), KEGG (Kanehisa, Furumichi, Sato, Ishiguro-Watanabe, & Tanabe, 2021) and MammaPrint (Cardoso et al., 2016), the transcription factors (TFs) documented in RegNetwork (Liu, Wu, Miao, & Wu, 2015) and the top-ranked feature genes from GCWs (Kong & Yu, 2018) are added into DEGs to derive a whole gene set. Thirdly, we extract the maximum connected component from an integrated gene interaction network obtained from RegNetwork. Thus, the corresponding gene expression profile on the largest connected subnetwork is obtained. Fourthly, we propose CNet-SVM model to perform the network-based feature selection with connectivity constraints. Fifthly, we identify connected network structural biomarker genes of BRCA and compare their performances with numerous feature selection-empowered SVM methods. Sixthly, we verify the effectiveness by function enrichment analysis and external independent datasets validation.

2.2. Support vector machine (SVM)

Considering the training dataset $\mathcal{D}$ including n samples on the space $\mathbb{R}^p \times \mathbb{R}$

$$\mathcal{D} = \{(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \ldots, (\boldsymbol{X}_n, y_n)\}, \tag{1}$$

where $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is the vector with p features of ith sample, $y_i \in \{-1, +1\}$ is the corresponding binary label for $\boldsymbol{X}_i$ with $i = 1, 2, \ldots, n$.

For classification, the aim of SVM is to find a linear hyperplane, i.e.,

$$f(\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\theta} + \theta_0 = \sum_{j=1}^{p} \theta_j x_{ij} + \theta_0 = 0, \tag{2}$$

such that it maximizes the distance between two classes (-1 and 1). In Eq. (2), $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top \in \mathbb{R}^p$ is a unknown p-dimensional vector of coefficients with $\|\boldsymbol{\theta}\|_2 = 1$ and $\theta_0 \in \mathbb{R}$ is the intercept.

Based on Eq. (2), for the test data, we can use $\text{sign}(f(\boldsymbol{X}_i^{new}))$ to predict the class of any new input sample $\boldsymbol{X}_i^{new}$, where $\text{sign}(\cdot)$ is a decision function:

$$\text{sign}(f(\boldsymbol{X}_i^{new})) = \begin{cases} 1, & f(\boldsymbol{X}_i^{new}) > 0, \\ -1, & f(\boldsymbol{X}_i^{new}) < 0. \end{cases} \tag{3}$$

SVM expects to find an optimal separating hyperplane (2) to reduce the misclassification (3) of new data as much as possible.

Suppose the high-dimensional data with p ($\gg n$) dimensional is linearly separable. We employ a linear SVM model for an a-proof-of-concept study. The linear function can be solved by the following soft margin SVM problem

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0} \quad & \tfrac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \mu \textstyle\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i \left(\boldsymbol{X}_i^\top \boldsymbol{\theta} + \theta_0\right) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n, \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{4}$$

where $\mu > 0$ is the tuning parameter that controls the balance between $\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$ (the coefficients of the hyperplane) and $\sum_{i=1}^{n} \xi_i$ (the upper bound of the empirical risk), and ξ_i is the slack variable denoting the difference of the ith sample to the required functional margin with

$$\xi_i : \begin{cases} = 0, & \text{if sample } i \text{ on the correct side of the margin,} \\ \in (0, 1], & \text{if sample } i \text{ inside the margin,} \\ \in (1, +\infty), & \text{if sample } i \text{ on the right side of the margin.} \end{cases} \tag{5}$$

2.3. Regularized support vector machine (Reg-SVM)

Similar to other regularized classification models (Li & Liu, 2020, 2021), Eq. (4) can be expressed in a more generalized form (loss function and regularization penalty term). Firstly, for two constraints in Eq. (4), they can be summarized as

$$\xi_i \geq \max\left\{0, 1 - y_i \left(\boldsymbol{X}_i^\top \boldsymbol{\theta} + \theta_0\right)\right\}, \quad i = 1, 2, \ldots, n. \tag{6}$$

Table 1
Different penalty functions for regularization terms commonly used in Eq. (8).

Method	Formula	Property	Reference
Lasso-SVM	$\mathscr{P}(\boldsymbol{\theta};\lambda)=\lambda\sum_{j=1}^{p}\lvert\theta_j\rvert$	Convex	Zhu et al. (2003)
ENet-SVM	$\mathscr{P}(\boldsymbol{\theta};\lambda)=\lambda\left[\alpha\sum_{j=1}^{p}\lvert\theta_j\rvert+(1-\alpha)\sum_{j=1}^{p}\theta_j^2\right],\quad \alpha\in(0,1)$	Convex	Zou and Hastie (2005)
SCAD-SVM	$\mathscr{P}(\boldsymbol{\theta};\lambda)=\sum_{j=1}^{p}\mathscr{P}_a\left(\lvert\theta_j\rvert;\ \lambda\right),$ where $\mathscr{P}_a(\lvert\theta\rvert;\lambda)=\begin{cases}\lambda\lvert\theta\rvert, & \lvert\theta\rvert\le\lambda,\\ \frac{-(\theta^2-2a\lambda\lvert\theta\rvert+\lambda^2)}{2(a-1)}, & \lambda<\lvert\theta\rvert\le a\lambda,\\ \frac{(a+1)\lambda^2}{2}, & \lvert\theta\rvert>a\lambda.\end{cases}$	Non-convex	Zhang et al. (2006)
L2SCAD-SVM	$\mathscr{P}(\boldsymbol{\theta};\lambda)=\lambda\left[\sum_{j=1}^{p}\mathscr{P}_a\left(\lvert\theta_j\rvert;\ \alpha\right)+(1-\alpha)\sum_{j=1}^{p}\theta_j^2\right]$	Non-convex	Becker et al. (2011)

Secondly, by defining the following function mapping relationship

$$[\triangle]_+=\begin{cases}\triangle, & \text{if } \triangle\ge 0,\\ 0, & \text{if } \triangle<0,\end{cases} \tag{7}$$

and combining with Eq. (6), then Eq. (4) can be rewritten as the following unconstrained optimization problem

$$\min_{\boldsymbol{\theta},\theta_0}\{\mathscr{L}(\boldsymbol{\theta}|\mathscr{D})+\mathscr{P}(\boldsymbol{\theta};\lambda)\}, \tag{8}$$

where $\mathscr{L}(\boldsymbol{\theta}|\mathscr{D})=\sum_{i=1}^{n}\left[1-y_i\left(\boldsymbol{X}_i^{\top}\boldsymbol{\theta}+\theta_0\right)\right]_+$ is the hinge loss function, $\mathscr{P}(\boldsymbol{\theta};\lambda)=\lambda\|\boldsymbol{\theta}\|_2^2$ is the regularization penalty term, where $\|\boldsymbol{\theta}\|_2^2=\sum_{j=1}^{p}\theta_j^2$ is the square of L_2-norm for coefficient and λ is the regularization parameter controlling the amount of penalization. In general, a larger λ would punish more coefficients θ_j $(j=1,2,\ldots,p)$ to zeros.

Eq. (8) is usually called regularized support vector machine (Reg-SVM) model. As shown in Table 1, the penalty function $\mathscr{P}(\boldsymbol{\theta};\lambda)$ can have different forms for specific purposes, where these four methods are embedded feature selection methods, which integrate feature selection with classifier training simultaneously (Li & Liu, 2020, 2021, 2022).

In contrast, the penalty terms of Lasso-SVM and ENet-SVM methods are convex, which ensure their local optimal solutions are globally optimal. While the penalty terms of SCAD-SVM and L2SCAD-SVM methods are non-convex, which have many good properties, such as sparsity (Xu, Zhang, Wang, Chang, & Liang, 2010), unbiased (Fan, Peng, et al., 2004), and oracle property (Knight, Fu, et al., 2000). Therefore, the SCAD method and its variants, e.g., L2SCAD, have received widespread attention recently. What is more, some models use more than one tuning parameter to balance the trade-off between loss function and model complexity. Without loss of generality, we set $\alpha=0.5$ for the alternative penalties in models. Considering the SCAD penalty is not sensitive in the selection of the tuning parameter a (Becker et al., 2011), we use the suggested value $a=3.7$ for SCAD-SVM and L2SCAD-SVM models.

Fig. 2 displays the one-dimensional images of the four penalty functions shown in Table 1. As shown in Fig. 2, when $\theta\in[-1,1]$, the imposed penalty effect of Lasso-SVM and SCAD-SVM is the same, and the penalty effect of ENet-SVM and L2SCAD-SVM is also the same. However, when $|\theta|\ge 1$, the penalty degrees of the four penalty terms are significantly different. As the coefficient increases gradually, the difference in the penalty effect becomes more and more apparent. If we impose a fixed penalty value, i.e., $\mathscr{P}(\theta)=2$, the increased order of regression coefficients θ is ENet-SVM, L2SCAD-SVM, Lasso-SVM and SCAD-SVM. That is to say, when the regression coefficient is greater than 1, the features selected by ENet-SVM are most sparse, while those selected by SCAD-SVM are least sparse theoretically. Moreover, the larger the θ is, the more obvious this trend will be.

2.4. Connected network-constrained SVM (CNet-SVM)

For performing feature selection in a network, we compile a gene–gene interaction network. Consider a prior gene network documented

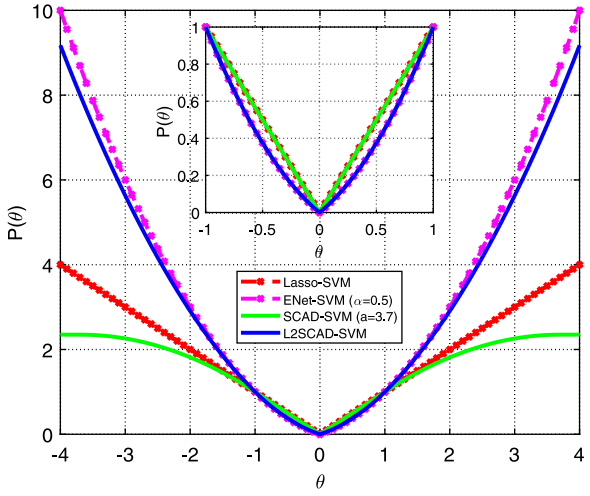


Fig. 2. The one-dimension images for the penalty functions shown in Table 1.

in RegNetwork (Liu et al., 2015) that is represented by a simple graph $G=(V,E)$ without self-loops or parallel edges, where $V=\{v_1,\ldots,v_p\}$ is a set of vertices mapped by genes, $E=\{e_{kj}\mid k\sim j\}$ is a set of edges indicating that gene k and j are linked in the knowledge-based gene network (Chen et al., 2011).

With the aim of imposing network connectivity in feature selection (Wang, Buchanan, & Butenko, 2017), we propose two definitions that kj-node cut $\mathcal{S}$ and kj-node cut set Γ at first. They are closely related to the connectivity constraints that we will impose on the Reg-SVM model. Considering that these two concepts have been introduced in detail in previous work (Li & Liu, 2022), we will not repeat them here. The basic definitions with corresponding mathematical representations can be found in Supplementary Materials Additional Information.

Suppose the subset C of graph G is composed of all vertices of the connected components of G, and let the subset U of C be a specific selected connected component. Here we only take care of the $\mathcal{S}$ when it is the minimal kj-node cut (Wang et al., 2017) and exemplify it by referring to Fig. 3. For instance, the set $\{4,5\}$ belongs to $\Gamma(3,6)$, but the set $\{2,4,5\}$ does not. According to the breadth-first search (BFS) algorithm (Trudeau, 2013), the minimal kj-node cut $\mathcal{S}$ of the studying component of the prior gene interaction network can be obtained.

In addition, Supplementary Materials Additional Information also give the definitions of normalized Laplacian matrix $\mathscr{L}$ of graph G, the *Dirac* measure $\delta_l(U)$ at a vertex v_l of set U and the theorem of modeling connectivity (Carvajal, Constantino, Goycoolea, Vielma, & Weintraub, 2013). Based on that, let $\mathscr{P}_\alpha(\boldsymbol{\theta})=\alpha\|\boldsymbol{\theta}\|_1+(1-\alpha)\boldsymbol{\theta}^{\top}\mathscr{L}\boldsymbol{\theta}$, then for $\forall\ k,j\in C,\ e_{kj}\notin E,\ \forall\ U\subseteq C\subseteq G$, we give the sparse and connected

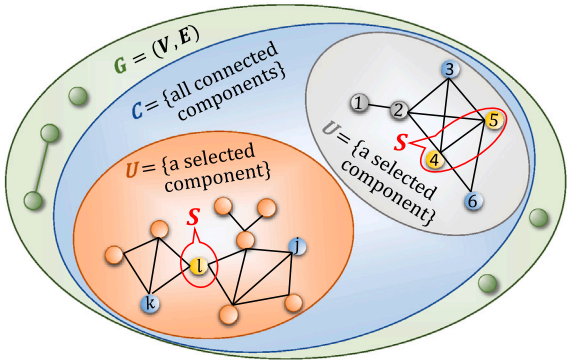


Fig. 3. The kj-node cut $\mathcal{S}$ in a selected connected component of a graph.

network-constrained (CNet) penalty as

$$\mathcal{P}(\boldsymbol{\theta};\lambda) = \lambda\mathcal{P}_{\alpha}(\boldsymbol{\theta}) = \lambda\alpha\|\boldsymbol{\theta}\|_1 + \lambda(1-\alpha)\boldsymbol{\theta}^{\top}\mathcal{L}\boldsymbol{\theta}, \tag{9}$$

with

$$\sum_{l\in\mathcal{S}} \boldsymbol{\delta}_l(U) \geqslant \boldsymbol{\delta}_k(U) + \boldsymbol{\delta}_j(U) - 1, \ \forall\ \mathcal{S} \in \Gamma(k,j), \tag{10}$$

where $\lambda > 0$ is the regularization parameter, $\alpha \in (0,1)$ is a mixing parameter that determines the ratio between sparse terms $\|\boldsymbol{\theta}\|_1$ and smooth penalty $\boldsymbol{\theta}^{\top}\mathcal{L}\boldsymbol{\theta}$, $\boldsymbol{\delta}_x(U)$ is 1 if $x \in U$ and 0 if $x \in C \setminus U$.

After including CNet penalty term (9) with the inequality constraints (10) into Eq. (8), we build up the novel CNet-SVM model as the following constraint programming:

$$\begin{aligned} &\min_{\boldsymbol{\theta},\theta_0} \ \sum_{i=1}^{n} \left[1 - y_i\left(\boldsymbol{X}_i^{\top}\boldsymbol{\theta} + \theta_0\right)\right]_+ + \lambda\alpha\|\boldsymbol{\theta}\|_1 + \lambda(1-\alpha)\boldsymbol{\theta}^{\top}\mathcal{L}\boldsymbol{\theta} \\ &\text{s.t.} \quad \boldsymbol{\delta}_k(U) + \boldsymbol{\delta}_j(U) - 1 \leq \textstyle\sum_{l\in\mathcal{S}} \boldsymbol{\delta}_l(U), \\ &\qquad \forall\ \mathcal{S} \in \Gamma(k,j),\ \forall\ k,j \in U,\ e_{kj} \notin E,\ \forall U \subseteq C \subseteq G. \end{aligned} \tag{11}$$

The convexity property of model (11) guarantees it has an optimal solution. With that in place, it can be solved by the interior-point method (Li & Liu, 2022).

2.5. Selection of parameters

Since the optimal tuning parameters are essential in performing classification and feature selection tasks, we provide search guidance on the regularization path in CNet-SVM model (11). Specifically, we investigate the behaviors of the parameter λ and α and achieve their theoretical property (Zhang et al., 2013).

Theorem 2.1. *Let* $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^{\top} \in \mathbb{R}^{n\times p}$, *and let*

$$\lambda_{\max} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_m\},$$

where $m \in \mathbb{N}^*$. *Given the dataset* $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$, *the maximum value of* λ *in model* (11) *with the solution* $\boldsymbol{\theta}^* \neq \mathbf{0}$ *satisfies*

$$\lambda_{\max} = \frac{1}{\alpha}\|\boldsymbol{X}^{\top}\boldsymbol{y}\|_{\infty}. \tag{12}$$

Theorem 2.1 shows that $\lambda_{\max}$ and α are inversely proportional, which plays a role in guiding the regularization path search. The proof of Theorem 2.1 can be available in the Supplementary Materials Additional Information.

In the experiments, we set different values for parameter α by dividing the interval (0, 1) into N equal parts. For each fixed α, we again take M different values for λ by equally splitting the interval $[1, \lambda_{\max}]$, then the parameter λ getting the minimum misclassification error can be picked up using the training data. For four Reg-SVM methods, let $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, the interval search algorithm (Frohlich & Zell, 2005) is used to optimize their parameter λ during the model training progress, respectively. In this work, the left and right endpoints of search interval are set as 2^{-10} and 2^{10} (Becker et al., 2011).

To investigate the feature selection behaviors of all Reg-SVM methods, we implement the K-fold cross-validation combined with search algorithms for tuning parameters. Namely, the training dataset is randomly split into K folds with equal sizes of samples. Meanwhile, each fold contains the corresponding responding variables of the dataset. We train the model on the $K-1$ folds data and test the prediction performances on the rest of one fold data. The procedure is performed K times to obtain the optimal parameter λ. Note that the choice of K affects the trade-off of variances and bias of the prediction error (Becker et al., 2011), and $K = 5$ has been recommended as a good compromise (Marcot & Hanea, 2021).

Table 2
The detailed information of the datasets used in this work.

Datasets	Platforms	# of genes	# of samples (Normal/BRCA)
TCGA	DCC	20501	1205 (112/1093[a])
GSE10780	GPL570	21835	184(143/41)
GSE10797	GPL571	13701	33(5/28)
GSE20437	GPL96	13749	42(42/0)
GSE21422	GPL570	21835	19(5/14)
GSE26910	GPL570	21835	12(6/6)
GSE31519	GPL96	13749	67(0/67)
GSE38959	GPL4133	19902	43(13/30)
GSE42568	GPL570	21835	121(17/104)
GSE45827	GPL570	15064	52(11/41)
GSE65194	GPL570	21835	66(11/55)

[a]Only 112 tumor samples that are from the same tissue origin and participants as 112 "Solid Tissue Normal" are extracted from 1093 "Primary Solid Tumor" samples to be used for biomarker discovery. For example, for a normal sample marked "TCGA-AC-A2FF-11A-13R-A17B-07", the case sample marked "TCGA-AC-A2FF-01A-11R-A17B-07" in tumor group will be extracted correspondingly.

2.6. Datasets and preprocessing

The primary BRCA dataset (discovery dataset) is downloaded from TCGA database. Meanwhile, the ten cohorts, compiled from GEO database including 639 samples (386 BRCA cases and 253 controls), are used as the external validation datasets to evaluate our discovered biomarkers. Table 2 lists the details of these datasets, the numbers in parentheses are the sample sizes. In this study, the BRCA dataset from TCGA consists of 112 adjacent normal breast tissues. In the actual experiments, for a fair contrast, we extract their corresponding 112 tumor samples. It ensures we choose the samples from the same tissue origin and participant. More importantly, a balanced dataset with the same number of positive and negative samples is established.

For data preprocessing, we first reduce the number of genes to 489 DEGs on the basis of adjusted P-value (P.adjust) and fold change (FC) by Wald test with Benjamini–Hochberg (BH) adjudication using DESeq (Love, Huber, & Anders, 2014). Then, combining 489 DEGs with 412 known BRCA-related disease genes or the genes with essential roles in the BRCA knowledgebase (82 elite cancer genes from MalaCards, 147 genes from KEGG BRCA pathway, 70 genes used by Mammaprint, 119 transcription factors (TFs) in RegNetwork and 169 top-rank genes from GCWs), we put these 901 genes into a prior human gene regulatory network documented in RegNetwork to compile them into an integrative gene interaction network. Finally, the maximum connected component consisting of 792 nodes and 10004 edges is obtained as the candidate feature network based on which we conduct network-based feature selection by CNet-SVM and other comparing methods. The full gene list can be available in Supplementary Materials Table S1.

Table 3
The classification results of selected features on the simulation data in case of $random.seed = 500$.

Method	SVM-RFE	mRMR-SVM	Lasso-SVM	ENet-SVM	SCAD-SVM	L2SCAD-SVM	GLasso-SVM	SGLasso-SVM	CNet-SVM
# of features	30	30	55	10	36	37	75	72	73
# of vertices	12	14	36	0	11	13	67	44	73
Ratio (%)	40.00	46.67	65.45	0	30.56	35.14	89.33	61.11	100
# of edges	7	9	31	0	8	9	77	38	82
Acc	0.778	0.639	0.639	0.778	0.639	0.611	0.750	0.778	0.889
Pre	0.765	0.600	0.591	0.800	0.591	0.565	0.750	0.714	0.842
Sn	0.765	0.706	0.765	0.706	0.765	0.765	0.706	0.882	0.941
Sp	0.789	0.579	0.526	0.842	0.526	0.474	0.789	0.684	0.842
F-measure	0.765	0.649	0.667	0.750	0.667	0.650	0.727	0.789	0.889
AUC	0.808	0.706	0.700	0.864	0.700	0.684	0.748	0.817	0.932

3. Results

3.1. On simulation data

In this section, we first conduct experiments to show our proposed CNet-SVM model's classification and prediction performance by several synthetic datasets. For a complete comparison study, we also explore the corresponding performances of one wrapper method: SVM-RFE (Guyon, Weston, Barnhill, & Vapnik, 2002), one filter method: mRMR-SVM (De Jay et al., 2013), four other embedded Reg-SVM methods: Lasso-SVM, ENet-SVM, SCAD-SVM and L2SCAD-SVM, and two grouping structure model: GLasso-SVM (Yang & Zou, 2015) and SGLasso-SVM (Huo et al., 2020). Specifically, we focus on their network connectivity structural behaviors in selecting features and their predictive precisions in internal validations.

3.1.1. Simulation scenario

First of all, we simulate the synthetic dataset $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ using the model

$$\boldsymbol{y} = 2 * (\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) - 3 * \mathbf{1}, \tag{13}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \{+1, -1\}$, $\mathbf{1} \in \mathbb{R}^n$ is a unit vector whose components are ones, $\boldsymbol{\theta} = [sample\{\mathcal{N}(0,1), 80\}, rep(0, p-80)]^\top$ is the coefficient vector with $\theta_0 = 0$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$ is the error term. The input matrix $\boldsymbol{X}$ representing gene expression data is created by $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is the covariance matrix with elements σ_{kj} equals to 0.6 if $1 \le k \ne j \le 80$, σ_{kj} is always be 1 when $k = j$ ($1 \le j \le 80$) and 0 otherwise. Thus it ensures the 80 features in matrix $\boldsymbol{X}$ are strongly correlated, i.e., there are more edge connections between genes to ensure the connectivity.

In addition, we construct the prior feature graph G represented by matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, where we connected the first 80 vertices with probability 3% to generate the major network structure and linked the rest vertices with probability 1%. In the experiment, we set sample $n = 120$, feature $p = 150$. By setting random seeds = {10, 30, 50, 100, 125, 160, 260, 300, 500, 700}, we generate ten different synthetic datasets.

3.1.2. Feature selection

For a case of the simulated data with $random.seed = 500$, we investigate the gene network and the features selected by CNet-SVM and the comparing eight methods. In contrast, because both SVM-RFE and mRMR-SVM realize feature selection by ranking the importance of features, we select the top 30 feature genes for subsequent analyses. The detailed results of feature selection, network structure and classification performance, e.g., accuracy (Acc), precision (Pre), sensitivity (Sn), specificity (Sp), F-measure and AUC are given in Table 3.

Table 3 shows the proposed CNet-SVM selects the most features and ENet-SVM selects the least. At the same time, their AUC values are ranked as first and second, respectively. Coincidentally, both Lasso-SVM and SCAD-SVM achieve the same AUCs. The AUCs of SGLasso-SVM, SVM-RFE, GLasso-SVM and mRMR-SVM, ranking the third, fourth, fifth and sixth, are only lower than those of ENet-SVM and CNet-SVM, respectively. While the L2SCAD-SVM obtains the lowest AUC value. Notably, from the network structure formed by selected features of these nine methods, the features identified by CNet-SVM form a completely connected network with 73 vertices and 82 edges, while the features from the ENet-SVM method are all isolated nodes. Regarding GLasso-SVM and SGLasso-SVM, the number of features selected by them is close to that selected by CNet-SVM, but the classification AUCs and network vertices ratios of which are significantly lower than CNet-SVM. It shows that CNet-SVM is superior in achieving higher classification performance. As expected, the selected features are from a connected network structure.

Moreover, the distribution between the features selected by these methods is visualized in Fig. 4, which implies the gene interactions, and this is closely related to the gene's cooperative functions. As implied in Fig. 4, there is a large amount of intersection between the features identified by CNet-SVM and those selected by the other methods, where we highlight the same feature nodes that appear in the other methods with squares. It can be seen that although the other eight methods also select the same features as CNet-SVM, many of them exist in isolation. In contrast, they are clearly connected in the feature network identified by CNet-SVM, which is consistent with the known candidate feature network information. On the other hand, we found that the three feature subnetworks identified by L2SCAD-SVM are all subnetworks in the feature network of CNet-SVM, and only one subnetwork identified by SVM-RFE or SCAD-SVM is not included in the feature network of CNet-SVM. This indicates that CNet-SVM not only ultimately maintains the original structure information between feature genes but also obtains better classification prediction results when performing feature selection.

3.1.3. Comparison studies

For the comparison study, we implement SVM-RFE, mRMR-SVM, four Reg-SVM methods, two grouping structural methods and CNet-SVM on ten synthetic datasets to explore the feature selection performances. In each synthetic data, 84 samples are randomly selected as the training data and the rest 36 samples are regarded as the test data, i.e., the internal dataset validation. For fairness, in each experiment, all nine models are conducted on the same training data and their performances in binary classification are computed in the same test data. The detailed results in the internal validations by the nine feature selection-empowered SVM methods in the ten experiments are given in the Supplementary Materials Table S2.

Without loss of generality, we calculate their average AUCs in the ten experiments to show their relative predictive performances. Fig. 5 demonstrates the comparison (average) AUCs on the ten simulation datasets, where mRMR-SVM obtains the lowest average AUC value and the AUCs of GLasso-SVM and SVM-RFE are slightly better than mRMR-SVM, but still inferior to the results of the four Reg-SVM methods with embedded feature selection and the SGLasso-SVM method. CNet-SVM achieves the highest AUCs in most cases, except for the 2nd (lower than

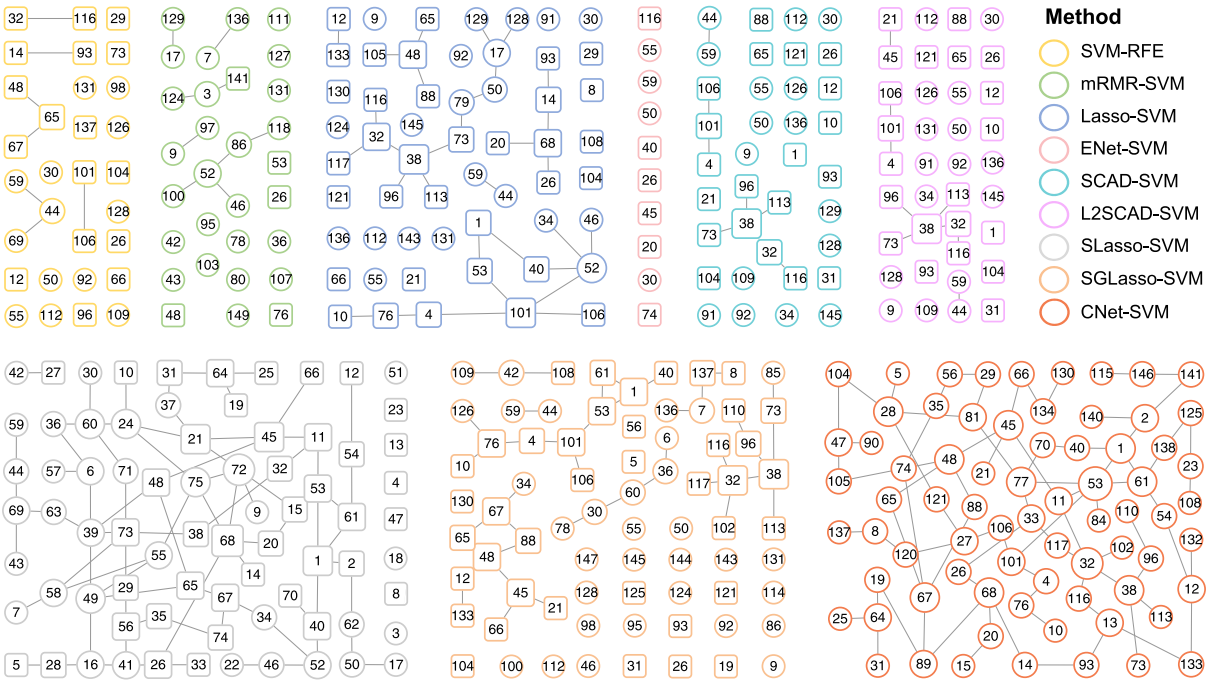


Fig. 4. The network-structured features selected by nine methods on simulated data with $random.seed = 500$, where different colors indicate different methods and the size of one node is proportional to its degree.

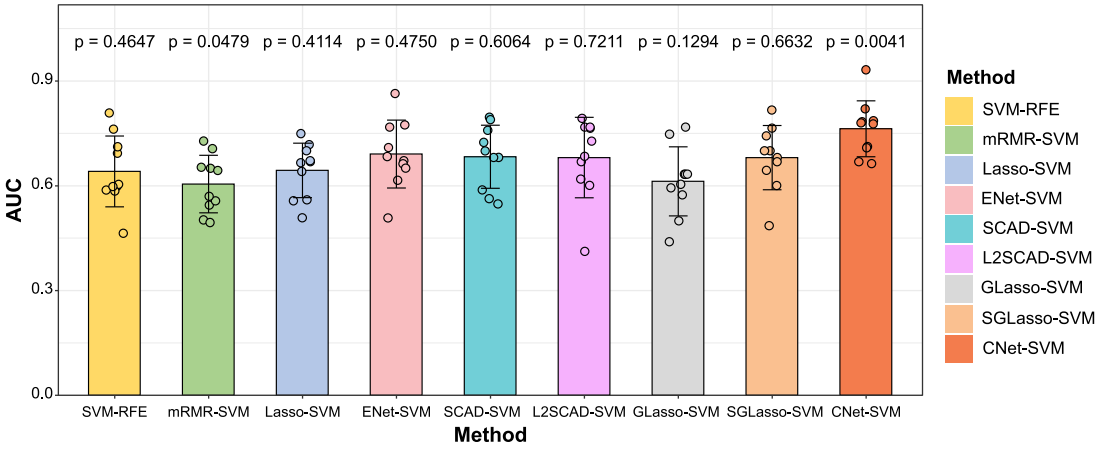


Fig. 5. The average AUC with error bar of nine comparing methods on ten different simulated datasets, where *P*-value refers to whether there is a significant difference within groups.

SCAD-SVM), 7th and 10th (both lower than L2SCAD-SVM) experiments. Obviously, from the whole level, the average AUC value (0.763 ± 0.076) of the CNet-SVM method is significantly higher than the other eight methods on the simulation data.

3.2. On real BRCA data

The research on the interactions between different genes helps get insight into biological interpretations of complex diseases such as cancer. In this section, to demonstrate the role of the proposed CNet-SVM method in network-based cancer biomarker discovery, we first apply it to identify network structural gene biomarkers of BRCA from the RNA-Seq data in TCGA. Based on that, we then analyze the enriched functions of identified biomarkers and validate their prediction accuracy on external independent datasets.

3.2.1. Network biomarkers discovery

CNet-SVM can successfully identify the known prior genes and unique genes to be explored. As shown in Fig. 6(a), our proposed CNet-SVM method identifies 32 genes, forming a connected network component with 36 edges without any redundant/isolated points. The feature network is a subnetwork of the original gene network. This reflects that CNet-SVM considers and preserves the structural information (i.e., interaction relationship) between genes during feature selection, which is crucial for discovering biologically meaningful biomarkers. Moreover, Fig. 6(a) also shows that among the biomarkers selected by CNet-SVM, there are 18 genes are differentially expressed in normal samples and tumor samples of BRCA, and eight genes are confirmed to be related to the BRCA pathway in KEGG, three genes are TFs in RegNetwork, and the remaining three genes belong to the specified ones identified by GCWs.

Using one wrapper method (RFE-SVM), one filter method (mRMR-SVM), four embedded Reg-SVM feature selection methods (Lasso-SVM,

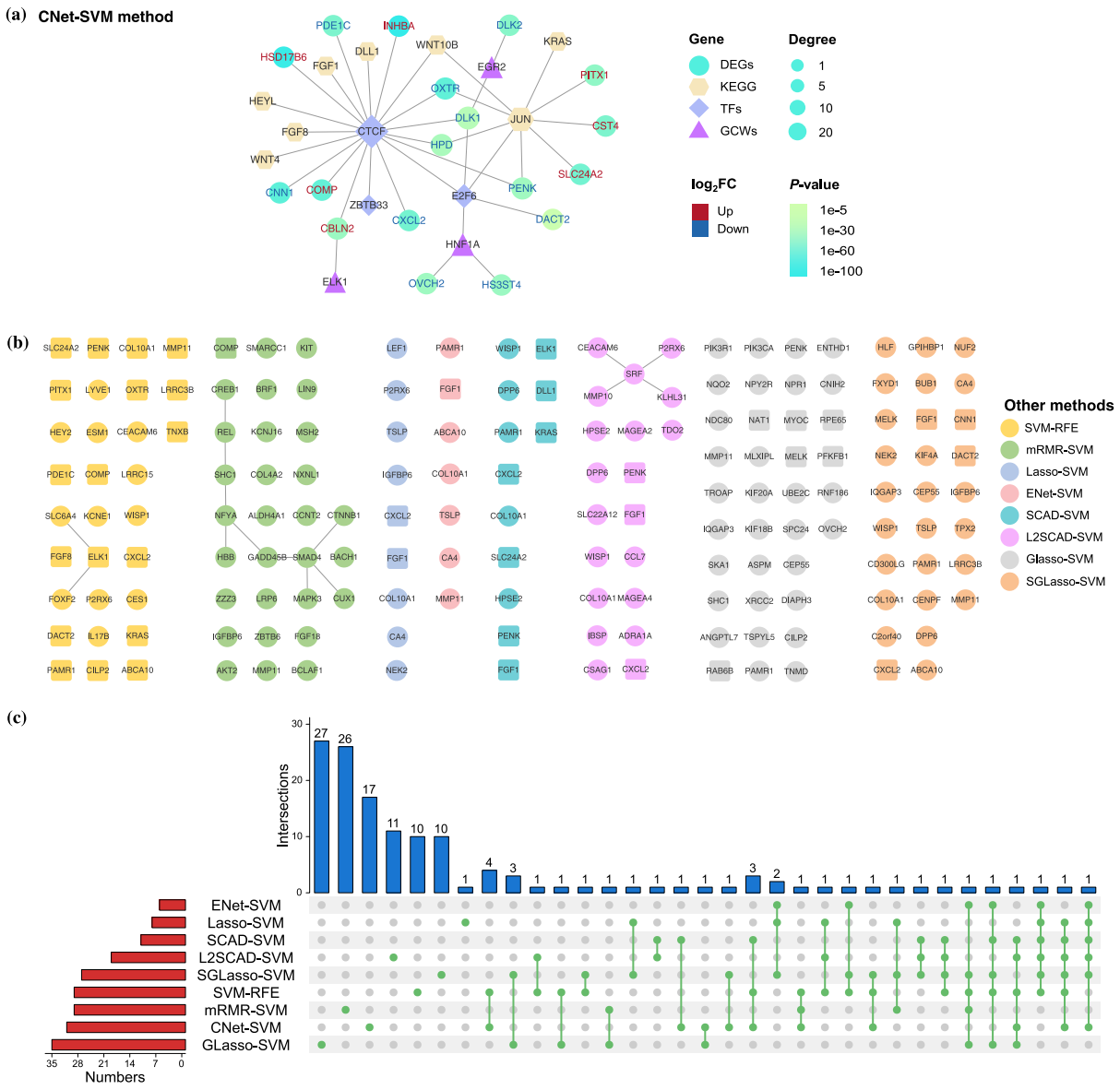


Fig. 6. The feature selection results of nine methods. (a) The network of biomarkers selected by our proposed CNet-SVM method, where the genes with red labels play the up-regulation roles and blue labels are down-regulations. (b) The features selected by SVM-RFE, mRMR-SVM, Lasso-SVM, ENet-SVM, SCAD-SVM, L2SCAD-SVM and SGLasso-SVM methods, where the square represents the gene that is also selected by the CNet-SVM method. (c) The overlap of features among nine methods, in which the top bar in blue exhibits the interactions among the methods, the overlaps are highlighted by the bottom dotted lines in green.

ENet-SVM, SCAD-SVM, L2SCAD-SVM) and two grouping structure methods (Glasso-SVM and SGLasso-SVM) on the prior feature network, we obtain eight feature gene subsets respectively. As conducted in the simulated data section, we also regard the top 30 features of SVM-RFE and mRMR-SVM as their selected features for subsequent analyses. In order to explore the network-structured features of identified biomarkers in the gene network, we visualize the genes picked by each method. As shown in Fig. 6(b), from the aspect of a network, Lasso-SVM, ENet-SVM and SCAD-SVM select fewer genes and these genes do not form into any network architecture but rather exist alone or isolated, without any connection to the gene network. GLasso-SVM and SGLasso-SVM select more genes, but all of them are isolated as well. In contrast, the results of SVM-RFE, mRMR-SVM and L2SCAD-SVM methods select more genes and some of them exhibit simple network structures. For instance, the selected feature genes of SVM-RFE and mRMR-SVM account for 1/10 and 1/3 of the prior gene network, respectively. L2SCAD-SVM selects 20 genes, but only 1/4 of them form a connected network with four edges.

We also analyze the overlaps between the genes in eight feature sets and the 32 biomarker genes to illustrate their similarities and differences. As shown in Fig. 6(c), among the 32 genes selected by CNet-SVM, 15 of them have intersections with the genes selected by the other eight methods. In addition, CNet-SVM also selected 17 unique genes that are not identified by other methods. The results imply that our method is similar to these existing methods but still maintains its unique preference in selecting feature genes. Last but not least, CNet-SVM preserves the initial structure between features, where all biomarker genes form a connected network component with the requested connectivity, indicating they cooperate with each other and perform physiological dysfunctions in BRCA.

3.2.2. *Function enrichment analysis*

In this section, to illustrate the functional implications underlying the biomarkers detected from CNet-SVM, we perform the functional enrichment analysis (Zhou et al., 2019). The enrichment of biological pathways helps find out how the molecules play a series of functions

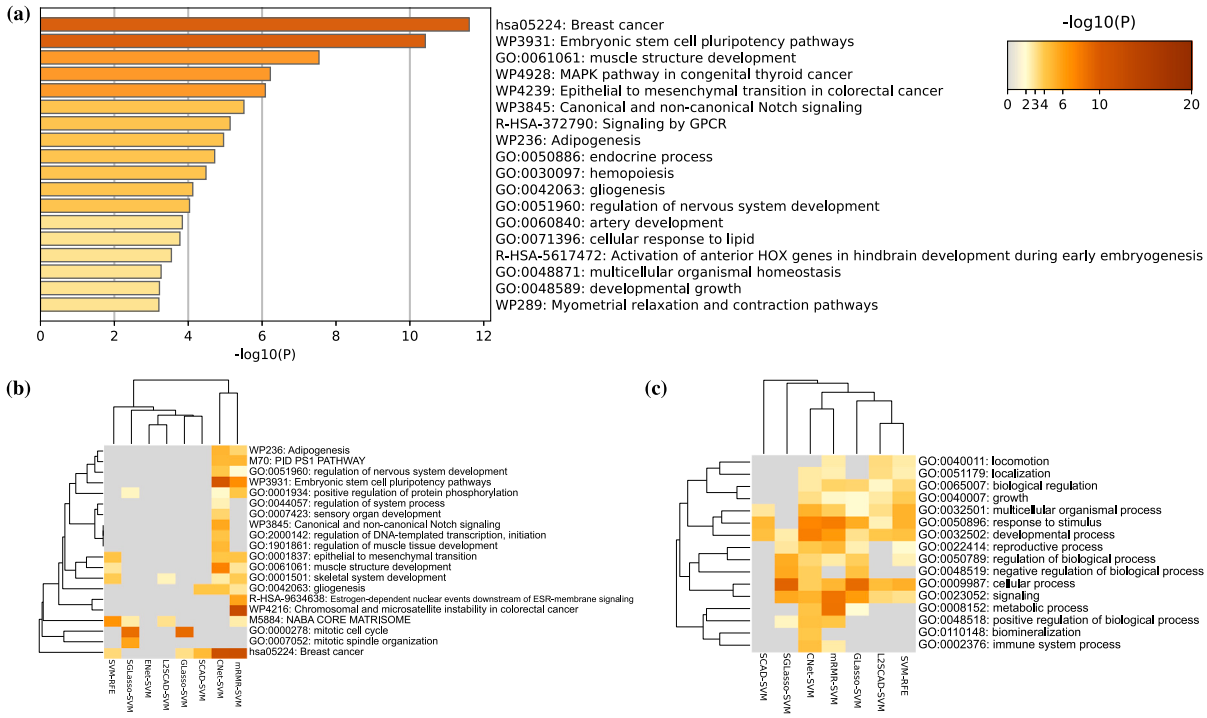


Fig. 7. The pathway and GO enrichment analysis of 32 identified biomarkers and the comparison results with the other methods. (a) The heatmap of the top 18 statistically enriched terms enriched by 32 biomarker genes selected by our proposed CNet-SVM method, where the accumulative hypergeometric test is used to calculate the *P*-value. The hierarchical tree clusters statistically enriched (b) pathway terms and (c) GO terms among different methods, where the gray cell indicates the lack of significance for the functional term in the corresponding method.

and lead to changes in a cell. Fig. 7(a) shows the top 18 representative terms with the best *P*-value enriched by our identified BRCA biomarkers via CNet-SVM, where the selected genes are most significantly subjected to "hsa05224: Breast cancer" pathway. In order to investigate the biological meaning, we show the top one enriched KEGG pathway with ID hsa05224 in Figure S1 in the Supplementary Materials Additional Information, where the red rectangular borders and fonts are our identified genes. Several terms in Fig. 7(a) are well-known for their association with BRCA and have been confirmed in experiments (Hanahan & Weinberg, 2000).

Besides, we also compare the enriched biological processes of gene subsets selected by other eight methods. Fig. 7(b) shows the top 20 enriched pathways and functional terms, wherein Lasso-SVM and ENet-SVM enrich no terms. SVM-RFE, SCAD-SVM and GLasso-SVM enrich some relevant terms and include the BRCA pathway, but this function term is not very significant. L2SCAD-SVM and SGLasso-SVM enrich some terms, but they do not cover the BRCA pathway. Clearly, the genes selected by CNet-SVM and mRMR-SVM are significantly enriched in the BRCA pathway.

Similarly, Fig. 7(c) shows the enriched gene ontology (GO) terms underlying the selected feature genes, in which both Lasso-SVM and ENet-SVM do not yield any results. While CNet-SVM obtains the most standardized annotation and interpretation on the functions of biomarkers, where GO:0050896 (response to stimulus), GO:0032502 (developmental process), GO:0032501 (multicellular organismal process) and GO:0048518 (positive regulation of biological process) have been validated by kinds of literature (Schvarcz et al., 2021; Shi et al., 2014; Smith et al., 2022; Zhao et al., 2021). Evidence shows that the four GO terms are highly relevant in human BRCA. In particular, we also select the enriched GO terms from the entire cluster enriched by the nine methods and convert them into a network layout. The comparing results are shown in Figure S2 in the Supplementary Materials Additional Information. These enriched results show that the biomarkers selected by CNet-SVM indeed imply the dysfunctional internal mechanism for BRCA. This approach is more efficient and effective in explaining the occurrence and development of BRCA as it involves structural information among gene–gene.

3.2.3. *Independent data validation*

CNet-SVM is a valuable and effective method for feature selection. We conduct external validations in independent data to verify the effectiveness and reliability of the BRCA biomarkers identified by CNet-SVM. For datasets GSE20437 and GSE31519 listed in Table 2, they only contain one phenotype, i.e., normal or tumor samples. Since they are from the same platform, we integrate them and regard these samples together as an independent dataset. Based on the 32 selected genes, we train an SVM classifier on the discovery data from TCGA and predict on the nine independent validation datasets, respectively. Then we obtain their classification AUCs as shown in Fig. 8. Intuitively, it shows that the biomarkers selected by CNet-SVM obtain higher prediction accuracy with AUC $>$ 0.950 on each external data, which displays the advantage of our proposed method in biomarker discovery by feature selection. The results of cross-dataset validations provide more evidence for the efficacy and advantage of CNet-SVM in identifying network-based biomarkers.

4. Discussion

Genes usually work collaboratively and many cancer-related genes orchestrate in an integrated pathway (Chen et al., 2011), they always perform dysfunctions in the form of a connected interaction network. However, the traditional Reg-SVM models assume that the phenotypic outcome is contributed by each feature individually (Chen et al., 2011). They conduct the embedded feature selection for classification by ignoring the connectivity characteristics even in so-called network-based methods. Therefore, the prediction performance and extension ability are often unexpectedly low when applying the identified features or biomarker genes in independent datasets, even if these validation data are obtained from similar genomic scenarios (Chen et al., 2011). From the perspective of interpretability, the connected network features

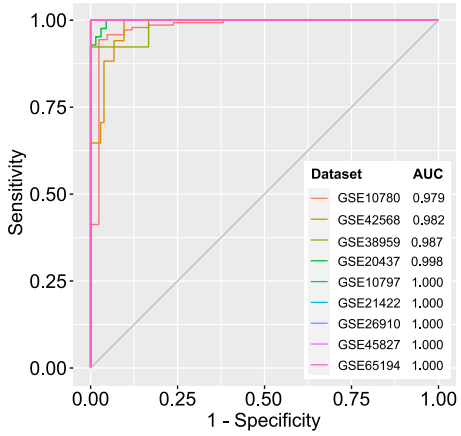


Fig. 8. The prediction results on nine external independent BRCA validation datasets of biomarkers discovered by CNet-SVM.

indicate the cooperative pathway of genes and products which are consistent with the actual condition in cells (Costanzo et al., 2019). From the perspective of functionality, the connectivity constraint in CNet-SVM makes the identified biomarkers directly form into a network enriched with easily interpretable functions (Wang et al., 2011).

Our proposed CNet-SVM method is an embedded feature selection in classification for identifying connected network-based biomarkers. Similar to the other Reg-SVM methods, we embedded the penalty function to achieve features' sparseness. This empowers the SVM model with feature selection ability. Unlike the other Reg-SVM methods, we introduced the network connectivity constraints into the model and formulated our model as a constrained optimization problem. The connectivity constraints ensure that the selected biomarkers are linked together and form into a connected network component. The other Reg-SVM models do not take the connectivity into consideration and just obtain several isolated features or some features that occasionally coexist in a small number of edges.

The numerous experiments conducted on simulation and real BRCA data demonstrate the effectiveness and efficiency of CNet-SVM. The comparison studies illustrate its superiority and advantage. CNet-SVM directly recognizes the prior network connectivity and effectively selects the network-based biomarkers consistent with the prior biological mechanism while maintaining the classification attributes simultaneously. The enriched functions underlying these interpretable biomarkers demonstrated that the network biomarkers have a closer functional relationship with BRCA. The external validations proved that these selected genes containing an inherent networking structure also perform better classification. For a proof-of-concept study, we studied the biomarker discovery in BRCA. The proposed CNet-SVM is flexible to be extended to discover diagnostic biomarkers for other complex diseases.

Currently, CNet-SVM aims to select network-based biomarkers to maintain the connectivity constraints. It is indeed helpful in interpreting the biological molecular mechanism underlying these selected biomarkers, and its results also outperform other alternative methods in terms of accuracy and interpretability. However, a noticeable request of our method is that the feature selection works on a prior gene network. The network structure heavily relies on our prior knowledge. That is to say, the network architecture is determined by knowledge-based gene circuits and pathways.

A complete and accurate map of prior network information is critical for network-based approaches. Recently, Kong et al. (2022) reported and discussed that the reduced predictive performance resulting from using an incomplete network highlights the importance of network coverage for identifying biomarkers. Thus, CNet-SVM might yield poor performance in some data where the networks are incomplete or unavailable. To alleviate the limitations and disadvantages derived from an incomplete network, on the one hand, we can integrate the prior gene interaction network (GIN) information from existing databases and literature, such as the network documented in our RegNetwork database (http://www.regnetworkweb.org) (Liu et al., 2015), the Transcription Factor Regulatory Network database (http://www.regulatorynetworks.org) (Neph et al., 2012) and many others (Chen & Liu, 2022). On the other hand, we have proposed a correlation-based reliable network construction framework when the prior GIN is unavailable (Liu, Liu, Zhao, & Chen, 2012). That is to say, we can build up a gene co-expression network before we implement the proposed network-based method. For the gene co-expression, we also have recognized its effectiveness in quantifying gene regulatory relationships with multiple association measures (Liu, 2017). The prior network structure in gene regulation and/or gene co-expression is a fundamental precondition for CNet-SVM.

Indeed, exploring a beneficial network interrelationship via prior knowledge plays a vital role in our follow-up research of the generalized CNet framework. We have presented some works for network construction. For example, we can use Pearson correlation, mutual information and Boolean threshold function to build a GIN from bulk or single-cell RNA-seq data. In this way, the proposed CNet-SVM model can be improved by exploring more credible prior network and pathway information. This is also essential to discover network biomarkers with better classification accuracy and feature interpretability.

5. Conclusions and future work

In this article, we developed a connected network-constrained SVM (CNet-SVM) method for biomarker discovery, which uses hinge loss as the objective function and considers the connectivity constraints between features. Firstly, we mathematically formulated CNet-SVM as a constrained optimization problem and theoretically provided search guidance on its regularization path. Secondly, we conducted two classes of experiments on several simulation data and real BRCA data to validate the classification and prediction performances of the CNet-SVM model. In particular, we compared CNet-SVM with one wrapper method, one filter method, four other embedded Reg-SVM methods and two grouping structure methods. The results showed that CNet-SVM performs better than the existing feature selection-empowered SVM methods. At last, we performed the functional enrichment analysis of identified biomarkers to indicate the disease mechanism related to BRCA and verified these biomarkers on nine independent datasets to illustrate that the biomarkers are extraordinarily effective and efficient.

Although the above encouraging experimental and comparable results indicate that the proposed method is indeed the right strategy to improve BRCA network-structured biomarker discovery, some aspects can be improved in future works. The main limitation of the proposed biomarker discovery technique is the generation of a reliable prior network or graph. The structural connections of prior networks are closely related to their underlying biological functions. In future works, one of our aims is to explore more credible a priori network information to decipher the more effective feature selection, revolutionize the more precise diagnosis, and produce more reliable and accurate cancer biomarker discovery. Another is to carry out the cell, animal and clinical practice based on the needs of translation medicine and use CNet-SVM and BRCA early risk screening technology to bring the discovered biomarkers into clinics. Our methods and findings are expected to provide more accurate early warning, diagnosis, and treatment of BRCA by translating the identified biomarkers into clinics.

CRediT authorship contribution statement

Lingyu Li: Conceptualization, Methodology, Data processing, Formal analysis, Software, Writing – review & editing. **Zhi-Ping Liu:** Investigation, Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under grant number 2020YFA0712402; National Natural Science Foundation of China (NSFC) under grant number 61973190; Natural Science Foundation of Shandong Province of China under grant number ZR2020ZD25 and Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) under grant number 2019JZZY010423; the Innovation Method Fund of China (Ministry of Science and Technology of China) under grant number 2018IM020200; the Scholarship under Shandong University's Exchange Program; the Fundamental Research Funds for the Central Universities, China under grant number 2022JC008; the Program of Qilu Young Scholars of Shandong University.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.120179.

Additional information. Some notation and definitions, the proof of Theorem 2.1 and two supplementary figures.

Table S1. The 489 DEGs and 412 known BRCA-related disease genes or the genes with essential roles in BRCA knowledgebase.

Table S2. The internal validation results of nine feature selection-empowered SVM methods in ten simulated datasets.

References

AAbraham, G., Kowalczyk, A., Zobel, J., & Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, *37*(2), 184–195. http://dx.doi.org/10.1002/gepi.21698.

Al-Obeidat, F., Tubaishat, A., Shah, B., Halim, Z., et al. (2020). Gene encoder: A feature selection technique through unsupervised deep learning-based clustering for large gene expression data. *Neural Computing and Applications*, 1–23. http://dx.doi.org/10.1007/s00521-020-05101-4.

Becker, N., Toedt, G., Lichter, P., & Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinformatics*, *12*(1), 1–13. http://dx.doi.org/10.1186/1471-2105-12-138.

Becker, N., Werft, W., Toedt, G., Lichter, P., & Benner, A. (2009). penalizedSVM: A R-package for feature selection SVM classification. *Bioinformatics*, *25*(13), 1711–1712. http://dx.doi.org/10.1093/bioinformatics/btp286.

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., et al. (2016). 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine*, *375*(8), 717–729. http://dx.doi.org/10.1056/NEJMoa1602253.

Carvajal, R., Constantino, M., Goycoolea, M., Vielma, J. P., & Weintraub, A. (2013). Imposing connectivity constraints in forest planning models. *Operations Research*, *61*(4), 824–836. http://dx.doi.org/10.1287/opre.2013.1183.

Chai, H., Huang, H., Jiang, H., Liang, Y., & Xia, L. (2016). Protein-protein interaction network construction for cancer using a new $L_{1/2}$-penalized net-SVM model. *Genetics and Molecular Research*, *15*(3), 1–14. http://dx.doi.org/10.4238/gmr.15038794.

Chen, G., & Liu, Z.-P. (2022). Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. *Bioinformatics*, *38*(19), 4522–4529. http://dx.doi.org/10.1093/bioinformatics/btac559.

Chen, L., Xuan, J., Riggins, R. B., Clarke, R., & Wang, Y. (2011). Identifying cancer biomarkers by network-constrained support vector machines. *BMC Systems Biology*, *5*(1), 1–20. http://dx.doi.org/10.1186/1752-0509-5-161.

Coleto-Alcudia, V., & Vega-Rodríguez, M. A. (2022). A multi-objective optimization approach for the identification of cancer biomarkers from RNA-seq data. *Expert Systems with Applications*, *193*, Article 116480. http://dx.doi.org/10.1016/j.eswa.2021.116480.

Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., et al. (2019). Global genetic networks and the genotype-to-phenotype relationship. *Cell*, *177*(1), 85–100. http://dx.doi.org/10.1016/j.cell.2019.01.033.

Cui, G., Fang, C., & Han, K. (2012). Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*, *13*(7), 1–10. http://dx.doi.org/10.1186/1471-2105-13-S7-S5.

De Jay, N., Papillon-Cavanagh, S., Olsen, C., El-Hachem, N., Bontempi, G., & Haibe-Kains, B. (2013). mRMRe: An R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, *29*(18), 2365–2368. http://dx.doi.org/10.1093/bioinformatics/btt383.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. http://dx.doi.org/10.1198/016214501753382273.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961. http://dx.doi.org/10.1214/009053604000000256.

Frohlich, H., & Zell, A. (2005). Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In *Proceedings. 2005 ieee international joint conference on neural networks, 2005, Vol. 3* (pp. 1431–1436). IEEE, http://dx.doi.org/10.1109/ijcnn.2005.1556085.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422. http://dx.doi.org/10.1023/A:1012487302797.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. http://dx.doi.org/10.1016/S0092-8674(00)81683-9.

Huo, Y., Xin, L., Kang, C., Wang, M., Ma, Q., & Yu, B. (2020). SGL-SVM: A novel method for tumor classification via support vector machine with sparse group lasso. *Journal of Theoretical Biology*, *486*, Article 110098. http://dx.doi.org/10.1016/j.jtbi.2019.110098.

Iqbal, S., & Halim, Z. (2020). Orienting conflicted graph edges using genetic algorithms to discover pathways in protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(5), 1970–1985. http://dx.doi.org/10.1109/TCBB.2020.2966703.

Jubair, S., Alkhateeb, A., Abou Tabl, A., Rueda, L., & Ngom, A. (2020). A novel approach to identify subtype-specific network biomarkers of breast cancer survivability. *Network Model Analaysis Health Information Bioinformation*, *9*(1), 1–12. http://dx.doi.org/10.1007/s13721-020-00249-4.

Jung, K.-M. (2013). Weighted support vector machines with the SCAD penalty. *Communications for Statistical Applications and Methods*, *20*(6), 481–490. http://dx.doi.org/10.5351/csam.2013.20.6.481.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551. http://dx.doi.org/10.1093/nar/gkaa970.

Kim, M., Park, J., Bouhaddou, M., Kim, K., Rojc, A., Modak, M., et al. (2021). A protein interaction landscape of breast cancer. *Science*, *374*(6563), 1–18. http://dx.doi.org/10.1126/science.abf3066.

Knight, K., Fu, W., et al. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378. http://dx.doi.org/10.1214/aos/1015957397.

Kong, J., Ha, D., Lee, J., Kim, I., Park, M., Im, S.-H., et al. (2022). Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nature Communications*, *13*(1), 1–15. http://dx.doi.org/10.1038/s41467-022-31535-6.

Kong, Y., & Yu, T. (2018). A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics*, *34*(21), 3727–3737. http://dx.doi.org/10.1093/bioinformatics/bty429.

Li, L., Ching, W.-K., & Liu, Z.-P. (2022). Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods. *Computational Biology and Chemistry*, *100*, Article 107747. http://dx.doi.org/10.1016/j.compbiolchem.2022.107747.

Li, L., & Liu, Z.-P. (2020). Biomarker discovery for predicting spontaneous preterm birth from gene expression data by regularized logistic regression. *Computational and Structural Biotechnology Journal*, *18*, 3434–3446. http://dx.doi.org/10.1016/j.csbj.2020.10.028.

Li, L., & Liu, Z.-P. (2021). Detecting prognostic biomarkers of breast cancer by regularized cox proportional hazards models. *Journal of Translational Medicine*, *19*(1), 1–20. http://dx.doi.org/10.1186/s12967-021-03180-y.

Li, L., & Liu, Z.-P. (2022). A connected network-regularized logistic regression model for feature selection. *Applied Intelligence*, *52*(10), 11672–11702. http://dx.doi.org/10.1007/s10489-021-02877-3.

Li, X., Liu, L., Goodall, G. J., Schreiber, A., Xu, T., Li, J., et al. (2020). A novel single-cell based method for breast cancer prognosis. *PLoS Computational Biology*, *16*(8), 1–20. http://dx.doi.org/10.1101/2020.04.26.062794.

Lin, S., Song, Q., Tao, H., Wang, W., Wan, W., Huang, J., et al. (2015). Rice_Phospho 1.0: A new rice-specific SVM predictor for protein phosphorylation sites. *Scientific Reports*, *5*(1), 1–9. http://dx.doi.org/10.1038/srep11940.

Liu, Z.-P. (2017). Quantifying gene regulatory relationships with association measures: A comparative study. *Frontiers in Genetics*, *8*, 96. http://dx.doi.org/10.3389/fgene.2017.00096.

Liu, X., Liu, Z.-P., Zhao, X.-M., & Chen, L. (2012). Identifying disease genes and module biomarkers by differential interactions. *Journal of the American Medical Informatics Association*, *19*(2), 241–248. http://dx.doi.org/10.1136/amiajnl-2011-000658.

Liu, Z.-P., Wu, C., Miao, H., & Wu, H. (2015). RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, *2015*, 1–12. http://dx.doi.org/10.1093/database/bav095.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. http://dx.doi.org/10.1186/s13059-014-0550-8.

Ma, S., Song, X., & Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, *8*(1), 1–17. http://dx.doi.org/10.1186/1471-2105-8-60.

Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, *36*(3), 2009–2031. http://dx.doi.org/10.1007/s00180-020-00999-9.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *70*(1), 53–71. http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., & Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, *150*(6), 1274–1286. http://dx.doi.org/10.1016/j.cell.2012.04.040.

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, *45*(D1), D877–D887. http://dx.doi.org/10.1093/nar/gkw1012.

Sarkar, J. P., Saha, I., Sarkar, A., & Maulik, U. (2021). Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific mirna biomarkers. *Computers in Biology and Medicine*, *131*, 1–13. http://dx.doi.org/10.1016/j.compbiomed.2021.104244.

Schvarcz, C. A., Danics, L., Krenács, T., Viana, P., Béres, R., Vancsik, T., et al. (2021). Modulated electro-hyperthermia induces a prominent local stress response and growth inhibition in mouse breast cancer isografts. *Cancers*, *13*(7), 1744. http://dx.doi.org/10.3390/cancers13071744.

Shi, W., Balazs, B., Györffy, B., Jiang, T., Symmans, W. F., Hatzis, C., et al. (2014). Combined analysis of gene expression, DNA copy number, and mutation profiling data to display biological process anomalies in individual breast cancers. *Breast Cancer Research and Treatment*, *144*(3), 561–568. http://dx.doi.org/10.1007/s10549-014-2904-z.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245. http://dx.doi.org/10.1080/10618600.2012.681250.

Smith, S., Stone, A., Oswalt, H., Vaughan, L., Ferdous, F., Scott, T., et al. (2022). Evaluation of early post-natal pig mammary gland development and human breast cancer gene expression. *Developmental Biology*, *481*, 95–103. http://dx.doi.org/10.1016/j.ydbio.2021.10.004.

Sun, Z., Fan, Y., Lelieveldt, B. P., & van de Giessen, M. (2015). Detection of Alzheimer's disease using group Lasso SVM-based region selection. In *Medical imaging 2015: computer-aided diagnosis, Vol. 9414* (pp. 285–291). SPIE, http://dx.doi.org/10.1117/12.2081368.

Tanvir, R. B., Aqila, T., Maharjan, M., Mamun, A. A., & Mondal, A. M. (2019). Graph theoretic and pearson correlation-based discovery of network biomarkers for cancer. *Data*, *4*(2), 81. http://dx.doi.org/10.3390/data4020081.

Trudeau, R. J. (2013). *Introduction to graph theory*. Courier Corporation.

Wan, S., Mak, M.-W., & Kung, S.-Y. (2012). mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics*, *13*(1), 1–16. http://dx.doi.org/10.1186/1471-2105-13-290.

Wan, S., Mak, M.-W., & Kung, S.-Y. (2016). Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins. *Journal of Theoretical Biology*, *398*, 32–42. http://dx.doi.org/10.1016/j.jtbi.2016.03.013.

Wang, Y., Buchanan, A., & Butenko, S. (2017). On imposing connectivity constraints in integer programs. *Applications of Management Science: In Productivity, Finance, and Operations*, *166*(1–2), 241–271. http://dx.doi.org/10.1007/s10107-017-1117-8.

Wang, J., Huang, Q., Liu, Z.-P., Wang, Y., Wu, L.-Y., Chen, L., et al. (2011). NOA: A novel network ontology analysis method. *Nucleic Acids Research*, *39*(13), e87–e98. http://dx.doi.org/10.1093/nar/gkr251.

Wang, H., Shao, Y., Zhou, S., Zhang, C., & Xiu, N. (2021). Support vector machine classifier via $L_{0/1}$ soft-margin loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12. http://dx.doi.org/10.1109/tpami.2021.3092177.

Wei, Y., Gu, F., & Zhang, W. (2021). A two-phase iterative machine learning method in identifying mechanical biomarkers of peripheral neuropathy. *Expert Systems with Applications*, *169*, Article 114333. http://dx.doi.org/10.1016/j.eswa.2020.114333.

Xu, Z., Zhang, H., Wang, Y., Chang, X., & Liang, Y. (2010). $L_{1/2}$ regularization. *Science China. Information Sciences*, *53*(6), 1159–1169. http://dx.doi.org/10.1007/s11432-010-0090-0.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *25*(6), 1129–1141. http://dx.doi.org/10.1007/s11222-014-9498-5.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *68*(1), 49–67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.

Zhang, H. H., Ahn, J., Lin, X., & Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, *22*(1), 88–95. http://dx.doi.org/10.1093/bioinformatics/bti736.

Zhang, W., Wan, Y.-w., Allen, G. I., Pang, K., Anderson, M. L., & Liu, Z. (2013). Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics*, *14*(S8), 1–8. http://dx.doi.org/10.1186/1471-2164-14-s8-s7.

Zhao, X., Guo, X., Jiao, D., Zhu, J., Xiao, H., Yang, Y., et al. (2021). Analysis of the expression profile of serum exosomal lncRNA in breast cancer patients. *Annals of Translational Medicine*, *9*(17), http://dx.doi.org/10.21037/atm-21-3483.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1), 1–10. http://dx.doi.org/10.1038/s41467-019-09234-6.

Zhu, J., Rosset, S., Tibshirani, R., & Hastie, T. J. (2003). 1-norm support vector machines. *Advances in Neural Information Processing Systems 10*, *16*(1), 16–23, https://proceedings.neurips.cc/paper/2003/file/49d4b2faeb4b7b9e745775793141e2b2-Paper.pdf.

Zhu, Y., Shen, X., & Pan, W. (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, *10*(1), 1–11. http://dx.doi.org/10.1186/1471-2105-10-s1-s21.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, H., & Yuan, M. (2008). The F_∞-norm support vector machine. *Statistica Sinica*, *18*, 379–398, https://www.jstor.org/stable/24308262.