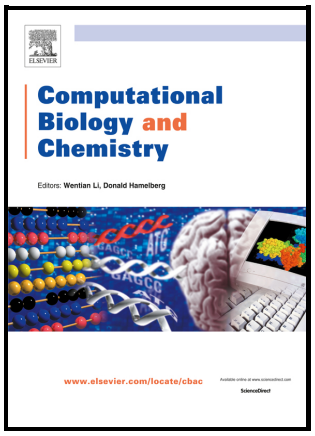
Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods

Lingyu Li, Wai-Ki Ching, Zhi-Ping Liu



PII: S1476-9271(22)00127-X

DOI: https://doi.org/10.1016/j.compbiolchem.2022.107747

Reference: CBAC107747

To appear in: *Computational Biology and Chemistry*

Received date: 14 April 2022
Revised date: 17 June 2022
Accepted date: 25 July 2022

Please cite this article as: Lingyu Li, Wai-Ki Ching and Zhi-Ping Liu, Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods, *Computational Biology and Chemistry,* (2022) doi:https://doi.org/10.1016/j.compbiolchem.2022.107747

Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods

Lingyu Li[a], Wai-Ki Ching[b], Zhi-Ping Liu[a,*]

[a]*School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China*
[b]*Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong*

Abstract

Recently, identifying robust biomarkers or signatures from gene expression profiling data has attracted much attention in computational biomedicine. The successful discovery of biomarkers for complex diseases such as spontaneous preterm birth (SPTB) and high-grade serous ovarian cancer (HGSOC) will be beneficial to reduce the risk of preterm birth and ovarian cancer among women for early detection and intervention. In this paper, we propose a stable machine learning-recursive feature elimination (StabML-RFE for short) strategy for screening robust biomarkers from high-throughput gene expression data. We employ eight popular machine learning methods, namely AdaBoost (AB), Decision Tree (DT), Gradient Boosted Decision Trees (GBDT), Naive Bayes (NB), Neural Network (NNET), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XGB), to train on all feature genes of training data, apply recursive feature elimination (RFE) to remove the least important features sequentially, and obtain eight gene subsets with feature importance ranking. Then we select the top-ranking features in each ranked subset as the optimal feature subset. We establish a stability metric aggregated with classification performance on test data to assess the robustness of the eight different feature selection techniques. Finally, StabML-RFE chooses the high-frequent features in the subsets of the combination with maximum stability value as robust biomarkers. Particularly, we verify the screened biomarkers not only via internal validation, functional enrichment analysis and literature check, but also via external validation on two real-world SPTB and HGSOC datasets respectively. Obviously, the proposed StabML-RFE biomarker discovery pipeline easily serves as a model for identifying diagnostic biomarkers for other complex diseases from omics data. The source code and data can be found at `https://github.com/zpliulab/StabML-RFE`.

Keywords: Robust biomarker discovery, Machine learning, Recursive feature elimination, Stable feature selection, Spontaneous preterm birth, High-grade serous ovarian cancer

1. Introduction

Exploring accurate and efficient disease biomarkers is an extremely challenging task in precision medicine (Abeel et al., 2010). Fortunately, high-throughput gene expression data provides new resources and opportunities. For instance, Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases provide a huge amount of data for discovery of biomarkers. Recently, more and more attention has been paid to the application of gene expression data in early complex disease diagnosis (Li and Guan, 2020). However, transcriptome usually uses a small number of samples to measure a large number of genes, which leads to great difficulties in selecting feature genes as biomarkers for the curse of dimensionality (Liu, 2016). Recursive feature elimination (RFE) is an effective alternative to evaluate feature importance and perform feature selection (Guyon et al., 2002). Thanks to the RFE technique, many machine learning methods are empowered for feature selection. It has implemented with many machine learning algorithms (such as AdBoost (AB) (Freund et al., 1996), Decision Tree (DT) (Quinlan, 1986), Gradient Boosted Decision Trees (GBDT) (Friedman, 2001), Naive Bayesian (NB) (Domingos and Pazzani, 1997), Neural Network (NNET) (Hecht-Nielsen, 1992), Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Sun et al., 2016) and XGBoost (XGB) (Chen and Guestrin, 2016)) to screen features, among

* Corresponding author
Email addresses: `lingyuli@mail.sdu.edu.cn` (Lingyu Li), `wching@hku.hk` (Wai-Ki Ching), `zpliu@sdu.edu.cn` (Zhi-Ping Liu)

which the least important genes are identified and then subsequently eliminated. The process is iterative until only one feature gene remains in feature-selection-based biomarker discovery (Sun et al., 2016).

RFE selects the optimal feature subsets and enhances the performance of classification by ranking through large number of features based on some specific machine learning method. However, the selected features tend to be unstable and un-robust and cannot be reproduced in other experiments (Bommert et al., 2022). Therefore, building up a stable feature selection process to screen robust biomarkers from gene expression data has become very important in various biomedical domains (Ge et al., 2019). The non-reproducibility of a list of selected feature genes is a topic of recent interest that many endeavors are trying to solve. Some strategy has been proposed, such as ensemble feature selection (EFS) that requires performing feature selection multiple times and aggregating the produced results (Guan et al., 2014; Chen et al., 2020; Mera-Gaona et al., 2021). Currently, there are three major types of ensemble techniques (i.e., data diversity, functional diversity and hybrid strategy) in the field of EFS from omics data. Specifically, data diversity uses the same feature selection method on different subsets of a dataset by repeated and random sampling (Awada et al., 2012). Functional diversity applies different feature selection techniques on the same dataset (Ben Brahim and Limam, 2018). In contrast, hybrid strategy conducts multiple feature selection techniques on multiple datasets (Chiew et al., 2019). A number of studies have suggested that EFS is a promising and cutting-edge approach to achieve suitable classification performances and stable features simultaneously.

The irreproducibility of biomarker identification remains one of the major obstacles to clinical applications (He and Yu, 2010). In other words, the robustness of gene selection is extremely important for biomarker discovery from gene expression data. To our best knowledge, few existing studies apply a stable ensemble feature selection technique that considers functional diversity to transcriptomic data. To date, biomarker identification mainly relies on the classification accuracy of the selected feature genes. For instance, Farinella et al. (2022) developed a computational pipeline to conduct machine learning and feature selection, where they used a two-step feature selection (correlation analysis and relief algorithm) and employed a DT classifier to find the diagnostic markers of HGSOC. Hwangbo et al. (2021) selected features using the stepwise selection method and constructed machine learning classifiers using logistic regression, RF, SVM and deep neural network to predict platinum sensitivity in HGSOC patients. Although these biomarker discovery methods have been available, they do not necessarily identify the stable feature subsets if we repeat the feature selection procedure. Even on the same training dataset, it is possible to find different feature subsets selected by the same method or by different ones that can still achieve competitive classification accuracy.

In this work, we develop a stable machine learning-recursive feature elimination pipeline called StabML-RFE, which provides a screening strategy by aggregating the classification performance based on AUC value and a stability metric based on Hamming distance for robust biomarker discovery from omics datasets. Firstly, eight ML-RFE methods (i.e., AB-RFE, DT-RFE, GBDT-RFE, NB-RFE, NNET-RFE, RF-RFE, SVM-RFE and XGB-RFE) are used to rank the importance of all features on the training data. Without loss of generality, we select the top-ranked genes as the optimal feature subset of each method. Then, we explore the classification performance of each optimal subset using a logistic regression classifier on the test data and select the subsets with a suitable AUC cut-off value. In particular, we use the Hamming distance to measure the stability corresponding to all combinations of the optimal feature subsets screened by AUC, and take the high-frequent genes in the combination with the maximum stability value as the output robust biomarkers. Finally, we conduct two case studies on SPTB and HGSOC datasets using our proposed StabML-RFE method. We also compare the selected biomarkers robustness with the other alternative ML-RFE methods. The internal validation and functional enrichment analysis provide a preliminary justification for the identified biomarkers. Moreover, we also verify them on external independent datasets and perform literature checks for further validation. All results demonstrate StabML-RFE is effective and efficient for screening robust biomarkers.

The rest of this paper is structured as follows. In Section 2, we give the general framework for discovering robust biomarkers from gene expression data. We propose the StabML-RFE method and its stable aggregation strategy. We further introduce a prediction risk score (PRS) measure for evaluating the disease risk. In Section 3, we assess the classification and prediction performances of our proposed method and compare it with numerous ML-RFE methods by conducting experiments on two real-world datasets (SPTB and HGSOC). Finally, we present the conclusions in Section 4 and discuss further research directions.

2. Materials and Methods

2.1. Datasets

The gene expression profiling datasets of SPTB and HGSOC are downloaded from NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). Table 1 lists the details of data, where the number in parentheses refers to the sample size. The discovery datasets are used to screen out robust biomarkers by the StabML-RFE method and the validation datasets are used to verify their effectiveness. Especially, for two datasets about SPTB, only the maternal whole blood samples are collected. While for datasets about HGSOC, because the samples of each one are small, we remove the genes in GSE27651 with no corresponding expression profiles in GSE69428, then unify the patients having the same measured genes as whole samples. Thus, the merged gene expression profiles consist of 48 patients and 17,689 genes. In addition, for each discovery dataset, we identify the differential expression genes (DEGs) by empirical Bayes method (Johnson et al., 2007) and adjust the P-values (denoted as P.adj) by the Benjamini and Hochberg (BH) method (Benjamini and Hochberg, 1995). The detailed DEGs list of SPTB and HGSOC are shown in the Supplementary Materials Table S1.

2.2. Framework

Fig. 1 illustrates the framework of StabML-RFE for discovering robust biomarkers from gene expression data. At first, we get the gene expression data from GEO database. On the discovery datasets, we identify DEGs with significant differences between control and case samples. We randomly divide them into the training data and test data with 7:3. Then, we implement the proposed method on the discovery datasets to identify robust biomarker genes. Specifically, StabML-RFE consists of three steps: First, we use eight machine learning methods with RFE to train and rank all the genes/features. Then the top-ranked α genes in the features ranked by each method are defined as the optimal feature subset obtained by the method, where $\alpha \in (0, 1)$ is a percentage parameter. Second, we train the logistic regression classifier using eight optimal feature subsets on the training data and verify their classification performance on the test data respectively. Subsequently, we establish a stability measure for all possible combinations of the optimal feature subsets whose AUC is larger than threshold τ to provide a stability/robustness evaluation of selected features, where parameter $\tau \in (0.5, 1)$ is determined by the predicted performance on the test dataset. Thirdly, in all possible combinations, the high-frequent features (greater than or equal to κ) corresponding to a maximum stability value will be screened as the output robust biomarkers, where parameter $\kappa \in [2, N-1]$ indicates that one feature is selected by how many methods. At last, we verify the efficacy of identified biomarkers by internal validations, functional enrichment analysis, literature check and external independent validations.

2.3. Machine learning classifiers

In this work, we simply focus on the binary classification scenario. Supposing there are n samples of observations independently and identically distributed with the training dataset

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

where $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is a p-dimensional gene expression data of the i-th sample, and $y_i \in \{0, 1\}$ is a label variable for $\boldsymbol{X}_i$ with $i = 1, 2, \cdots, n$.

There are numerous popular machine learning algorithms for binary classification. Here we use the following eight methods, namely AdBoost (AB), Decision Tree (DT), Gradient Boosted Decision Trees (GBDT), Naive Bayesian (NB), Neural Network (NNET), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XGB), to learn on the discovery data. Considering that they are all very well-known, we will not repeatedly describe them in this section and only present the brief introductions. The detailed information of each classifier including the chosen parameters can be found in the Supplementary Materials Methodology.

In classification, DT, NB, NNET and SVM are categorized as eager learners. They construct classification models based on training data before receiving new data for prediction. For training dataset $\mathcal{D}$, the DT classifier recursively partitions the feature space with vector $\boldsymbol{X}_i \in \mathbb{R}^p$ and label vector y such that the samples with the same label values are grouped together (Quinlan, 1986). Given data $\boldsymbol{X}_i$, the NB model predicts the class C_k for $\boldsymbol{X}_i$ according to the probability $p(C_k|\boldsymbol{X}_i) = p(C_k|x_{i1}, x_{i2}, \cdots, x_{ip})$, $\forall\ k \in 1, \cdots, K$, where K is all possible outcomes. Specifically, for the given the class C_k, the NB model assumes that feature x_{il} is independent of feature x_{ij} for $l \neq j$ (Domingos and Pazzani, 1997). For the NNET classifier, the formal neuron has p inputs $x_{i1}, x_{i2}, \cdots, x_{ip}$ that model the signals coming from dendrites, where the inputs are labeled with the corresponding synaptic weights $w_1, w_2, \cdots, w_p$ that measure their permeabilities. After reaching fixed threshold, the value of excitation level ξ induces the output y of the neuron, where the non-linear

Table 1
The detailed information of datasets used for robust biomarkers discovery of SPTB and HGSOC.

Diseases	Functions	Datasets	Platforms	# of samples	# of genes	Phenotypes of samples	References
SPTB	For discovery (326)	GSE59491	GPL18694	326	24,478	Term (228) / SPTB (98)	(Heng et al., 2016)
	For validation (17)	GSE73658	GPL1657	17	20,909	Term (12) / SPTB (5)	(Bukowski et al., 2017)
HGSOC	For discovery (48)	GSE69428	GPL570	20	17,689	Normal (10) / HGSOC (10)	(Yamamoto et al., 2016)
		GSE27651	GPL570	28	21,999	Normal (6) / HGSOC (22)	(King et al., 2011)
	For validation (271)	GSE120196	GPL570	14	21,999	Normal (4) / HGSOC (10)	(Au-Yeung et al., 2020)
		GSE14407	GPL570	24	21,999	Normal (12) / HGSOC (12)	(Bowen et al., 2009)
		GSE26712	GPL570	195	13,749	Normal (10) / HGSOC (185)	(Vathipadiekal et al., 2015)
		GSE40595	GPL570	38	21,999	Normal (6) / HGSOC (32)	(Yeung et al., 2013)

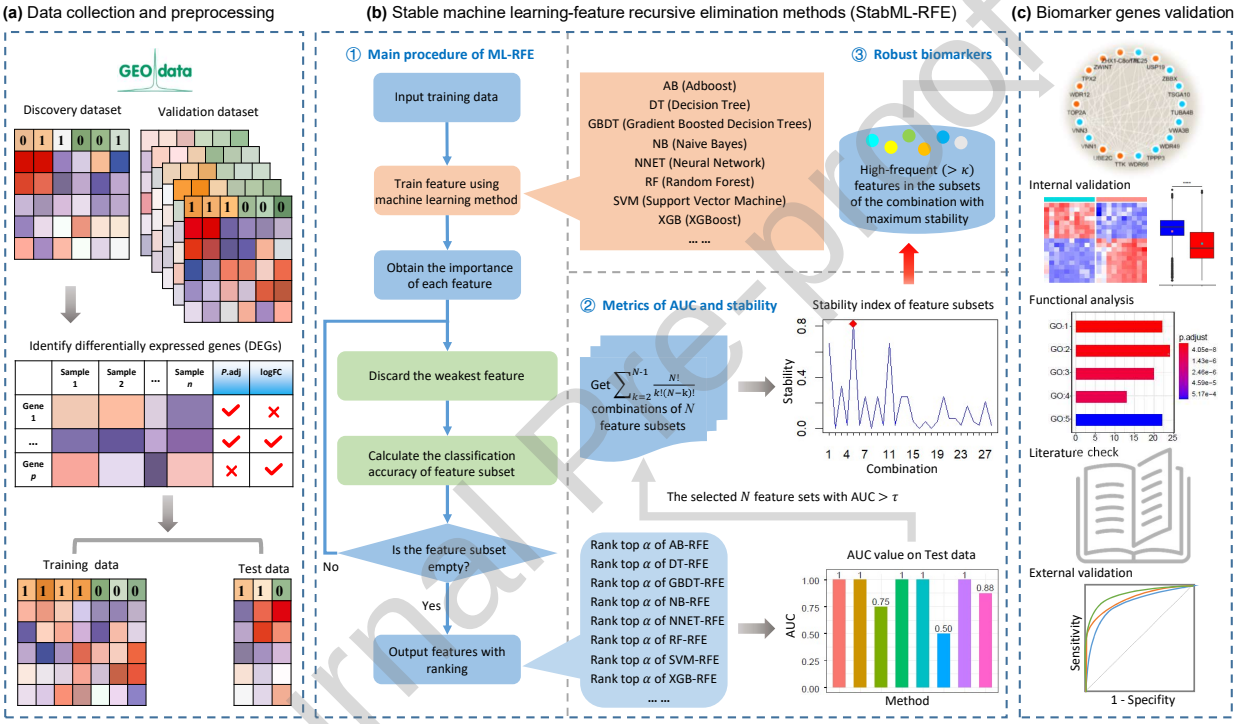


Fig. 1 The framework of StabML-RFE method for screening robust biomarkers from gene expression data.

grow of output $y = \sigma(x)$ is determined by the activation function σ (Hecht-Nielsen, 1992). Briefly, SVM is a classifier defined in the feature space to maximize the interval between two classes. When the data is linearly separable, SVM finds the optimal classification hyperplane $\omega^{\mathrm{T}} \cdot \boldsymbol{X}_i + b = 0$ and separates the samples of the two classes completely (Cortes and Vapnik, 1995).

In contrast, the other four machine learning algorithms, AB, GBDT, RF and XGB, are ensemble methods. The goal of AB is to identify the class that generated a particular instance by inferring over a collection of labeled observations provided as tuples $(\boldsymbol{X}_i, y_i)$ (Freund et al., 1996). The purpose of GBDT is to find a model F to fit the data such that the predicted value $\hat{y}_i$ for the j-th training example $\boldsymbol{X}_i$ is approximately equal to the j-th target value y_i or equivalently $\hat{y}_i = F(\boldsymbol{X}_i) \sim y_i$ for $\forall\ i = 1, 2, \cdots, n$ (Friedman, 2001). The RF model constructs many individual decision trees and pools the predictions from all trees to make the final prediction of the classes, where the importance of a node is calculated using Gini importance. The final feature importance of RF is the average over all the trees (Breiman, 2001). Finally, the XGB model is a scalable machine learning system for tree boosting. Considering that it is impossible to enumerate all the possible tree structures, the XGB model uses weighted quantile sketch to decide the candidate split points so as to calculate the total loss after split minus the total loss before split (Chen and Guestrin, 2016).

2.4. Recursive feature elimination (RFE)

Considering the huge amount of features of gene expression profiles on the training data, it is easy to cause over-fitting for a machine learning method due to the high dimensionality. This will result in low classification accuracy and weak generalization ability (Guyon et al., 2002). In fact, not all features or attributes or genes have positive effects on the prediction. Recursive feature elimination (RFE for short) is an effective and efficient technique for reducing the model complexity by removing irrelevant predictors. Technically, RFE is a wrapper feature selection algorithm that also uses filter feature selection internally, therefore different machine learning algorithms can easily combine with RFE in the core to perform feature selection (Guyon et al., 2002). As shown by the main procedure of ML-RFE in Fig. 1(b), it works by firstly fitting the machine learning model using all the features in the training set, then ranking features by importance (coefficient or feature importance), discarding the weakest feature(s) progressively one by one, and re-fitting the model. This process is repeated until a specified number of features is reached.

2.5. Stable machine learning-recursive feature elimination (StabML-RFE)

The robustness of biomarkers is an extremely important issue in candidate discovery because they greatly influence subsequent biological validations and clinical applications (Abeel et al., 2010). However, for the eight ML-RFE methods, they may yield different feature subsets although they are conducted on the same dataset. Therefore we propose a stability-based ensemble feature selection method called StabML-RFE to determine how much variation there is in the distribution of features which are selected subsets by different methods and to screen out the accurate and robust features. Fig. 2 illustrates the ensemble feature selection process based on the stable feature aggregation strategy in StabML-RFE.

Firstly, we use the former eight machine learning classifiers as the core of RFE to conduct the classification and feature selection simultaneously. After implementing the eight ML-RFE methods (AB-RFE, DT-RFE, GBDT-RFE, NB-RFE, MMET-RFE, RF-RFE, SVM-RFE and XGB-RFE) on the training data $\mathcal{D}$ with n sample and p features, we obtain eight feature subsets

$$\mathcal{FS} = \{FS_1, FS_2, \cdots FS_8\}, \tag{2}$$

where each feature subset contains p selected features

$$FS_i = \{f_{i1}, f_{i2}, \cdots, f_{ip}\}, \quad i = 1, 2, \cdots, 8, \tag{3}$$

with the corresponding rank

$$Rank_i = \{r_{i1} = 1, r_{i2} = 2, \cdots, r_{ip} = p\}. \tag{4}$$

Then, we determine the cut-off points for the features ranked by eight ML-RFE methods. Specifically, in the ranking of all features, we select the feature whose rank is greater than or equal to $\alpha \in (0, 1)$. Generally speaking, these features will be regarded as the most important features. In this way, $\lfloor \alpha p \rfloor$ features in each feature subset are selected to construct the individual optimal feature subset

$$FS_i^{opt} = \{f_{i1}, f_{i2}, \cdots, f_{i,\lfloor \alpha p \rfloor}\}, \quad i = 1, 2, \cdots, 8, \tag{5}$$

where $\lfloor \cdot \rfloor$ represents the round-down operator in mathematics.

After that, with the direction of AUC value on the test data and the stability metric of Hamming distance, we integrate all the optimal feature subsets by stable feature aggregation to screen out a final accurate and robust features. Namely, suppose that there are N optimal feature sets whose predictive classification performance AUC is large than the given parameter τ, and they are marked as

$$\mathcal{FS}^{opt} = \{FS_1^{opt}, FS_2^{opt}, \cdots FS_N^{opt}\}, \quad 3 \le N \le 8. \tag{6}$$

Equivalently, we consider the robust biomarker screening issue as a stable combination problem of N different feature subsets. We compute the stability of all possible combinations cases of the sets in $\mathcal{FS}^{opt}$. From the combination of any two subsets C_N^2, the possible combination of feature subset to C_N^{N-1} according to the number of feature subsets selected by the eight ML-RFE methods. Thus totally there are $\sum_{k=2}^{N-1} \frac{N!}{k!(N-k)!}$ possible combinations for $N \ge 3$.

Finally, the combination of $\mathcal{FS}^{opt}$ getting the maximum stability value is indicative to identify the final target feature set. We select the features in the combinations with maximum stability value by the principle of who appears more frequently (larger and equal to the fixed parameter κ) as the screened robust biomarkers,

$$\begin{aligned} \mathcal{FS}_{Stab}^{opt} = \Big\{ f_{ij}^{\kappa} \,|\, & f_{ij} \in \text{Combination}_m, \\ & \text{Frequency}(f_{ij}) \ge \kappa, \\ & \text{Stability}|_{\text{Combination}_m} = \text{maxmum.} \Big\}, \end{aligned} \tag{7}$$

where $1 \le i \le N$, $1 \le j < \lfloor \alpha p \rfloor$ and $1 \le m \le \sum_{k=2}^{N-1} \frac{N!}{k!(N-k)!}$.

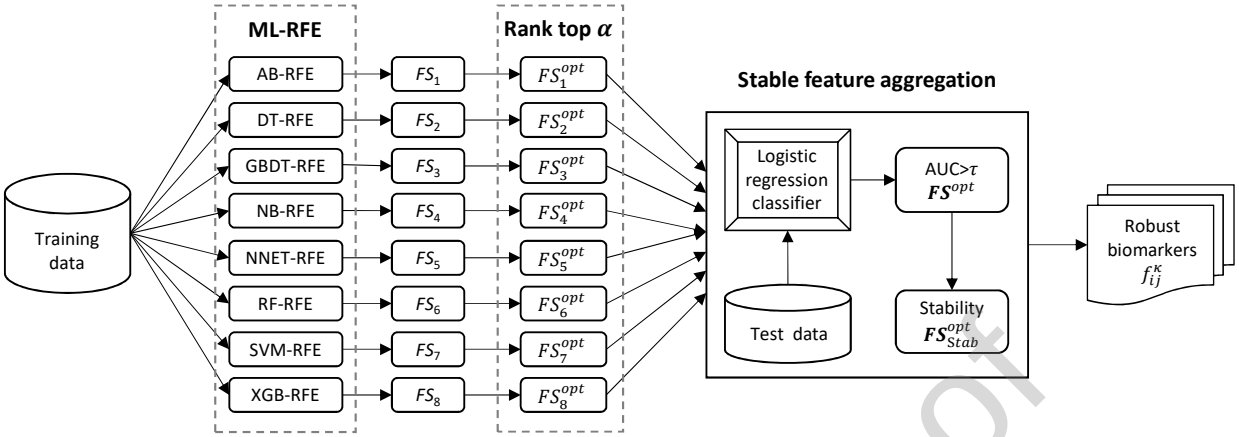


Fig. 2 The ensemble framework of StabML-RFE by stable feature aggregation.

2.6. Stability metric

The stability of feature selection is defined as the robustness of the set of selected features with respect to different methods (Zhang and Liu, 2021). Let $V_1, \cdots, V_k$ denote the k feature subsets of k feature selection methods, V_i^c denotes complement of set V_i for $i = 1, 2, \cdots, k$, and $|V|$ denotes the cardinality of certain a set V to be studied (Bommert and Lang, 2021). Also, let p denotes the total number of features in k feature sets. The stability of the feature selection is assessed based on the sets of selected features using the average normalized Hamming stability measure (Dunne et al., 2002), which is defined as

$$\text{Stability} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\left|V_i \cap V_j\right| + \left|V_i^c \cap V_j^c\right|}{p}, \tag{8}$$

where $\frac{|V_i \cap V_j| + |V_i^c \cap V_j^c|}{p}$ measures the similarity of the two sets V_i and V_j, which considers features that are included in both sets ($|V_i \cap V_j|$) as well as features that are not included in both sets ($|V_i^c \cap V_j^c|$). The maximum value of the stability measure is 1 that indicates a perfectly stable feature selection (Bommert and Lang, 2021).

2.7. Prediction risk score (PRS)

For each observation from the given dataset $\mathcal{D}$, the logistic regression model describes the relationship of binary outcome y_i and sample $\boldsymbol{X}_i$ by a Bernoulli distribution with the probability (Li and Liu, 2022),

$$p(\boldsymbol{X}_i) = \Pr(y_i = 1 \mid \boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}{1 + \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}, \tag{9}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$. Here θ_j $(j = 1, 2, \cdots p)$ is the coefficients to be estimated, θ_0 is the intercept.

Applying the logit transformation on Equation (9), it derives

$$\begin{aligned} \text{logit}(p(\boldsymbol{X}_i)) &= \log \frac{p(\boldsymbol{X}_i)}{1 - p(\boldsymbol{X}_i)} \\ &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip}. \end{aligned} \tag{10}$$

Based on the logistic regression classifier for training and validation given by Equation (10), we define the indicator of prediction risk score (PRS) for quantifying the disease risk of i-th sample to be observed (Li and Liu, 2020).

$$\text{PRS}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_q x_{iq}, \tag{11}$$

where q is the number of screened biomarkers by the proposed StabML-RFE method.

3. Results

In this section, we will demonstrate the effectiveness and efficiency of the features selected by the StabML-RFE method on two real-world datasets in SPTB and HSOC respectively. We also aim to confirm its accurate and robust performances in selecting features that contain important and concise information of classifying samples. For justifying our findings of biomarkers, we perform the differential expression analysis, clustering analysis, correlation analysis, as well as co-expression analysis. Moreover, we conduct gene ontology (GO) functional enrichment analysis and carry out necessary literature verification for the identified biomarkers. For external validation purpose, we also verify the identified

biomarkers on independent datasets for sample classification. In our comparison study, we compare our proposed StabML-RFE method with the other comparable ML-RFE methods to provide direct evidence of the effectiveness of StabML-RFE in discovering diagnostic biomarker genes.

3.1. On SPTB data

For discovering SPTB biomarkers, we first investigate the classification performance of eight optimal feature subsets with the parameter $\alpha = 0.045$ using an logistic regression classifier on the test data, the results in terms of accuracy (Acc), precision (Pre), sensitivity (Sn), specificity (Sp), F-measure and AUC are presented in Table 2. Clearly, from the aspect of AUC value, AB-RFE, GBDT-RFE, NNET-RFE and SVM-RFE methods obtain the relatively higher AUC values under the parameter $\tau = 0.800$, therefore these four optimal subsets are selected to construct set $\mathcal{FS}^{opt}$ to make the subsequent stability analysis.

Then we use the stability measure defined in Equation (8) to evaluate the stability of different optimal feature subsets in FS_i^{opt}, $i = 1, 2, 3, 4$. Intuitively, Figure 3(a) gives the curve of stability values under different combinations of $\mathcal{FS}^{opt}$, where the red diamond refers to its maximum point. The numbers in the x-axis coordinate represent all combinations of four optimal feature subsets selected by AB-RFE, GBDT-RFE, NNET-RFE and SVM-RFE, respectively. As known, a large stability value is desirable because it indicates a consistent choice of features in different machine learning with feature selection methods. Fig. 3(a) shows that the eighth combination induced of AB-RFE, GBDT-RFE and NNET-RFE achieve the maximum stability value, so the combination can be chosen to select robust features. Correspondingly, Fig. 3(b) shows the overlap genes selected by these three methods, where the genes colored with green means the selected features. Finally, we screen out the 24 high-frequent genes guided by $\kappa = 2$ as the robust biomarkers of SPTB.

Next, we explore the differential gene expression of the identified 24 robust biomarkers on the discovery data and show them in Fig. 4(a). It can be found that these genes illustrate significant differences in SPTB samples compared with those in term-birth (Term) samples, where the P-value is calculated using Welch's t-test and coded by symbols indicating statistical significance. We further implement the functional enrichment analysis using Metascape database (Zhou et al., 2019) to investigate the underlying functions. The results are shown in Fig. 4(b). As shown, SPTB is related to metabolic process, cellular process and signaling, and more detailed results are available in the Supplementary Materials Table S2. To further justify the reliability of screened biomarkers, we also make external validation using a logistic regression classifier on the independent SPTB dataset GSE73685. Fig. 4(c) shows the prediction and classification performance of robust biomarkers compared with the other optimal feature subsets selected by the three ML-RFE methods that used in stability aggregation processing. Clearly, StabML-RFE obtains the most promising classification result with AUC= 0.783. Although the AUC value of the NNET-RFE method is also competitive, it uses more features. After all, AB-RFE, GBDT-RFE and MMET-RFE selected 31 features, respectively. It illustrates the biomarkers obtained by the StabML-RFE reach better prediction performance than the other methods.

In the case study of SPTB, we further investigate the existing literature and match the identified 24 biomarkers with the prior-known biomarkers identified by other studies. There are eight overlap genes (FADS2, GLYR1, MFSD4A, PLEC, PRKAG1, GOLGA7, VAMP2 and ZNF284) that have been discovered by Li and Liu (2020) and nine overlap genes (MFSD4A, PLEC, VAMP2, PRKAG1, GLYR1, ACADVL, VNN1, ANKRD54 and PKHD1L1) of them are also reported by Rasmussen et al. (2022). Besides, some of the remaining 12 genes have also been confirmed to be associated with SPTB. For example, Plunkett et al. (2010) found that PLA2G4C influences preterm birth risk by increasing levels of prostaglandins. Yamada et al. (2019) reported a 31 years old pregnant woman who delivered a term baby via cesarean section but suffered from muscle weakness soon after delivery. Eventually she was identified as a homozygous p.K382Q mutation in ACADVL. These results prove the efficiency and creditability of the screened biomarkers by StabML-RFE.

3.2. On HGSOC data

For HGSOC biomarkers, we conduct StabML-RFE on the merged discovery dataset to obtain eight optimal feature subsets $\mathcal{FS}^{opt}$ using parameter $\alpha = 0.050$ in eight ML-RFE methods. For each one, we apply the logistic regression classifier on the training dataset and evaluate their prediction classification performances on the test data. Finally, we found that the six optimal feature subsets selected from AB-RFE, GBDT-RFE, NNET-RFE, RF-RFE, SVM-RFE and XGB-RFE methods achieve higher AUC values. Under the direction of parameter $\tau = 0.980$, six optimal subsets are taken as set $\mathcal{FS}^{opt}$ to make further analysis.

In order to obtain the maker genes with robustness, we extract all the combinations (a total of 56 combi-

Table 2
The comparison of classification performances using different ML-RFE methods on the test data of SPTB.

Performances	AB-RFE	DT-RFE	GBDT-RFE	NB-RFE	NNET-RFE	RF-RFE	SVM-RFE	XGB-RFE
Acc	0.784	0.711	0.753	0.722	0.794	0.649	0.773	0.742
Pre	0.643	0.522	0.581	0.545	0.655	0.419	0.621	0.567
Sn	0.621	0.414	0.621	0.414	0.655	0.448	0.621	0.586
Sp	0.853	0.838	0.809	0.853	0.853	0.735	0.838	0.809
F-measure	0.632	0.462	0.600	0.471	0.655	0.433	0.621	0.576
AUC	**0.850**	0.630	**0.800**	0.754	**0.885**	0.598	**0.819**	0.737

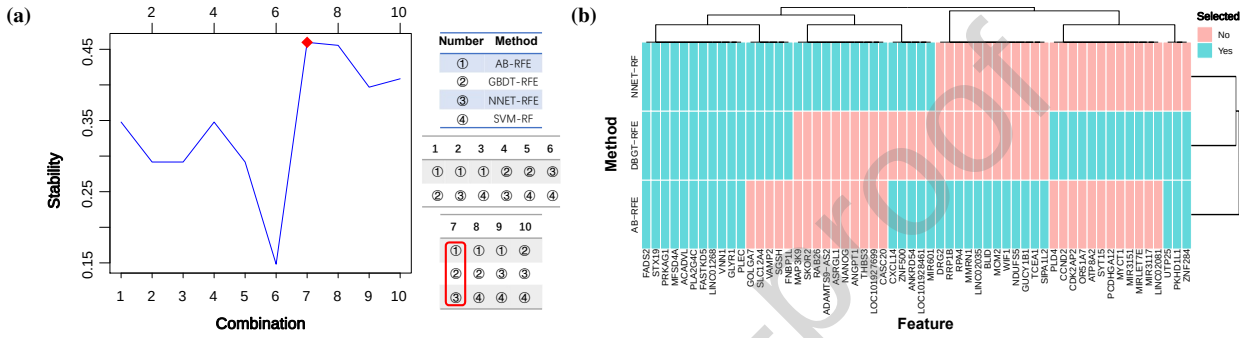


Fig. 3 Robust SPTB biomarkers identified by StabML-RFE. (a) Hamming stability change curve in different combinations. (b) The feature subset corresponding to the eighth combination and the selected robust biomarkers.

nations here) of $\mathcal{FS}^{opt}$ and calculate the stability value under each combination. Fig. 5(a) assesses the feature selection stability of the six ML-RFE methods, where the one-to-one correspondences and combinations of the numbers in the x-axis can be found in the Supplementary Materials Table S3. It is found that the 53-th combination (consisting of five methods: AB-RFE, GBDT-RFE, NNET-RFE, SVM-RFE and XGB-RFE) obtains the maximum stability value. Thus the 14 genes with high frequency with $\kappa = 2$ in the combination are regarded as the output robust biomarkers. Instinctively, Fig. 5(b) shows the detailed interactions (overlapping status) among these five individually feature subsets. It can be seen that AB-RFE, GBDT-RFE, NNET-RFE, SVM-RFE and XGB-RFE pick out more than 14 unique genes, while only one gene can be selected by at most four methods among them. This indicates that there are certain instability and diversity between these selected features although all of them achieve high classification and prediction performance.

Then, to validate the effectiveness and efficiency of our discovered HGSOC biomarkers, we explore their expression characteristics from the perspective of differential gene expression in the discovery data. The results are shown in Fig. 6(a), where the degree of significance is measured by Welch's T-test for each biomarker gene between the HGSOC and Normal groups. As shown, the biomarker genes illustrate significant difference in normal samples compared with those in HGSOC samples. Among them, the expressions of CRABP2, CENPF, KRT7 and VGLL1 are significantly down-regulated, and the rest are up-regulated. We also make the cluster analysis between these 14 biomarkers, Fig. 6(b) demonstrates the gene expression heatmap of identified 14 biomarkers of HGSOC. The whole samples are distinctively divided into two categories with a ratio 2:1, which is completely consistent with the real labels.

Enrichment analysis is also conducted to reveal the biological processes and dysfunctions associated with biomarkers for interpreting the underlying mechanisms of HGSOC. In the identified 14 biomarkers, we perform the functional enrichment analysis using Matascape as that in SPTB. Fig. 7(a) shows the enriched GO terms of biological process. In details, these HGSOC biomarkers are mainly related to metabolic process, detoxification, response to stimulus and biological regulation. They are consistent with the prior knowledge of tumorigenesis about HGSOC. For instance, Nath et al. (2021) proposes the potential of new therapeutic avenues that specifically target the metabolic pathways to overcome the chemoresistant of HGSOC. This verifies the biomarkers from the functional perspective and in turn proves the effectiveness of our proposed ensemble feature se-

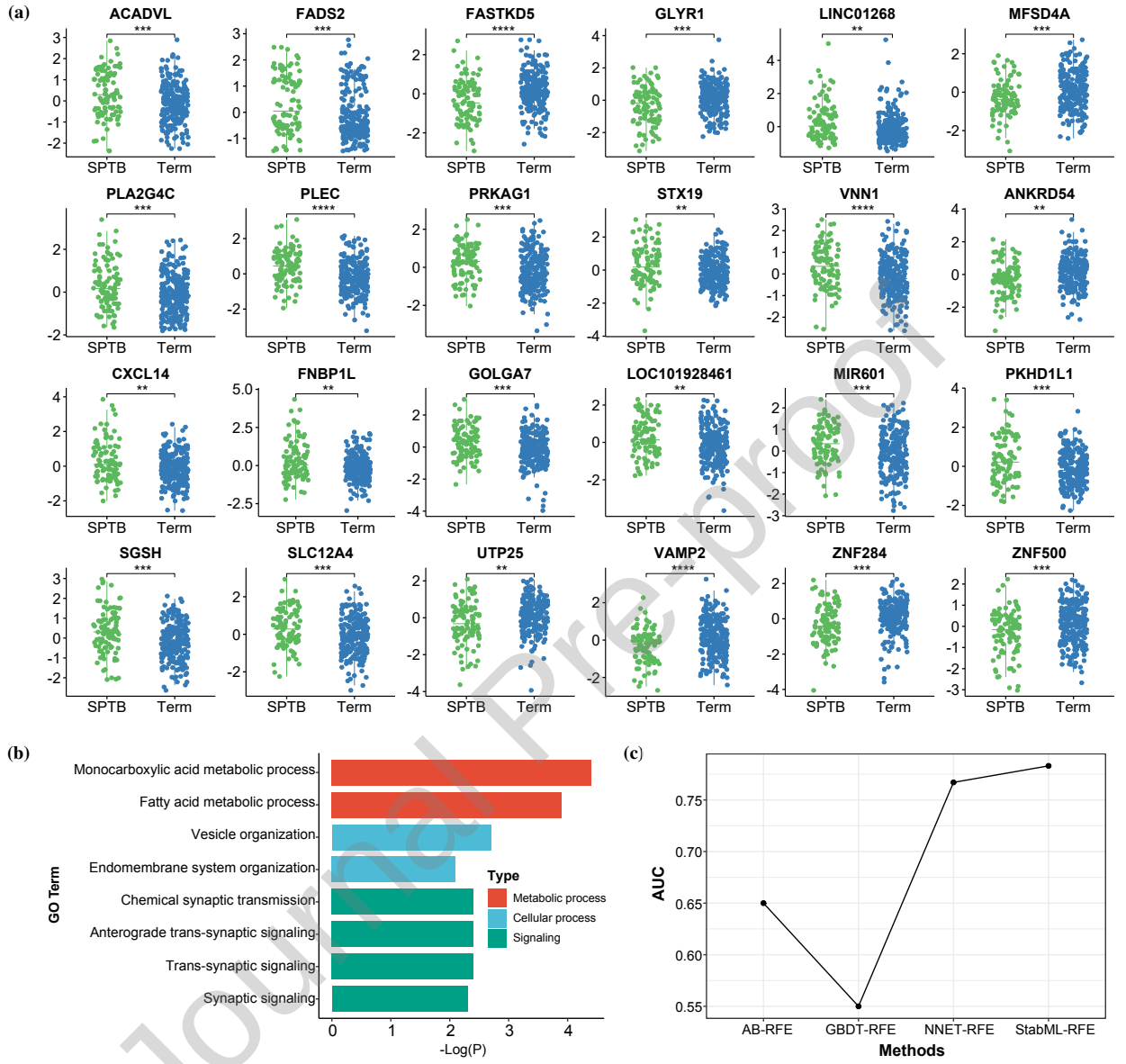


Fig. 4 The validation of 20 SPTB biomarkers screened by StabML-RFE. (a) The gene expression differential information on the GSE59491 dataset, where the significance is calculated by Welch's t-test with '*': $p < 0.05$, '**': $p < 0.01$, '***': $p < 0.001$ and '****': $p < 0.0001$. (b) The enriched GO terms. (c) The predictive AUC value on independent GSE73685 dataset.

lection method for biomarker discovery.

In addition, we make the correlation analysis between these 14 biomarkers to investigate their gene co-expression status. Fig. 7(b) not only presents the lower triangular matrix drawn by Pearson's correlation coefficients but also shows the co-expressed network architecture based on higher correlations, where 14 genes are connected by 44 edges with an absolute value of coefficient greater than 0.600. Apparently, there are co-expression relationships between these identified biomarkers, which indicates the identified biomarkers are not isolated but with close relationships with each other. Based on that, we further perform the network ontology analysis (Wang et al., 2011) of the biological network. The results can be obtained in the Supplementary Materials Table S4, which greatly enriches the net-

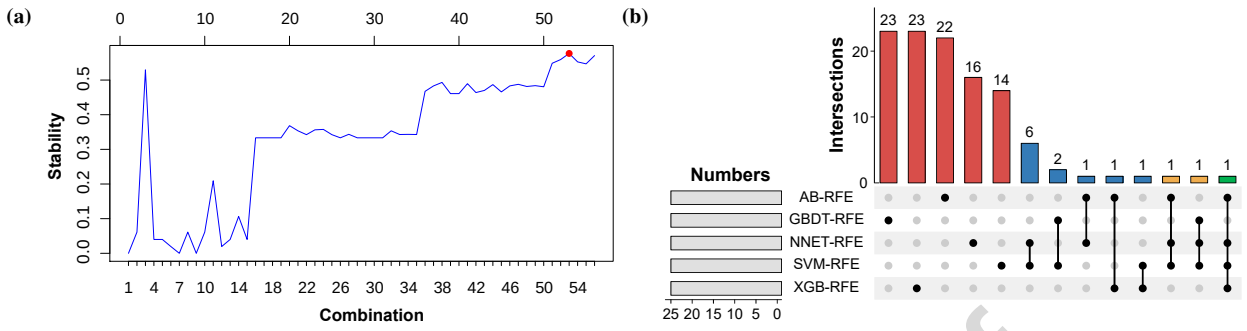


Fig. 5 Robust HGSOC biomarkers identified by StabML-RFE. (a) Hamming stability change curve in different combinations. (b) The overlaps of the optimal feature sets contained in the combination with the highest stability value.

work biological significance of biomarkers.

Furthermore, to further justify the screened 14 robust biomarkers of HGSOC, we successively perform four independent external data validations to comprehensively reveal the reliability and credibility of these biomarkers. In total, 271 samples have been tested. In particular, it is pointed out that we no longer consider three methods that perform poorly in both AUC and stability, and only compare with the other five methods that contribute to the stable feature aggregation strategy. For the fairness of comparison, we select the same number genes as the biomarkers identified by StabML-RFE, i.e., 14 top-ranked features in the five individual feature subsets $\mathcal{FS}^{opt}$. In each validation dataset, we also employ the evaluation metric of AUC to assess their classification performance. Table 3 illustrates the AUC values obtained from StabML-RFE and the comparing five ML-RFE methods. The average AUC value of StabML-RFE reaches 0.981 with a standard deviation of 0.019. Especially, it is worth noting that the screened biomarkers achieve very accurate classification and prediction on the dataset of GSE14407. These results prove the efficiency and advantage of identified biomarkers in classifying HGSOC samples from controls.

At last, in order to further confirm the validity of the 14 robust biomarkers, we verify them by checking the existing literature. The information reported by literature can be found in the Supplementary Materials Table S5, where eight genes have been confirmed to be closely related to the occurrence and development of HGSOC. These results prove the efficiency and creditability of the screened biomarkers by StabML-RFE. Although the remaining six genes have not been confirmed so far, they are likely to be used as the potential biomarkers need to be further validated.

3.3. PRS verification

Additionally, we calculate the PRS index for both SPTB and HGSOC on the independent validation datasets, and it can be used as a screening score to identify high-risk samples. Figure 8 (a) shows the violin plot of PRS for SPTB on one independent dataset. Figure 8 (b) shows the violin plot of PRS for HGSOC on four independent datasets. It is clear that there is a significant difference in the PRS indices for the disease samples when compared to the normal samples (P-value is calculated by Wilcoxon test). Therefore, PRS provides a useful approach to identifying samples at high risk of disease from a huge number of people. It encourages a person with a higher risk value to perform further test intervention for the early accurate diagnosis.

3.4. Robustness comparison

Essentially, the proposed StabML-RFE method is an ensemble feature selection method for biomarker discovery. Multiple optimal feature subsets are aggregated with the direction of AUC values and stability indices in order to increase the robustness of the final selected features. In this section, we will measure the robustness of screened signatures (i.e. biomarker genes) using the stability measure defined in Equation (8). Based on the different optimal feature subsets screened by computational methods in terms of AUC value and stability index, we respectively investigate the stability index of the discovered biomarkers by our proposed StabML-RFE method and compare with the features selected by other methods. The comparison results are shown in Fig. 9, where the upper triangular matrix and the down triangular matrix represent the stability value of SPTB and HGSOC biomarkers, respectively.

As shown in Fig. 9, for SPTB, the four methods, DT-RFE, NB-RFE, RF-RFE and XGB-RFE, are excluded

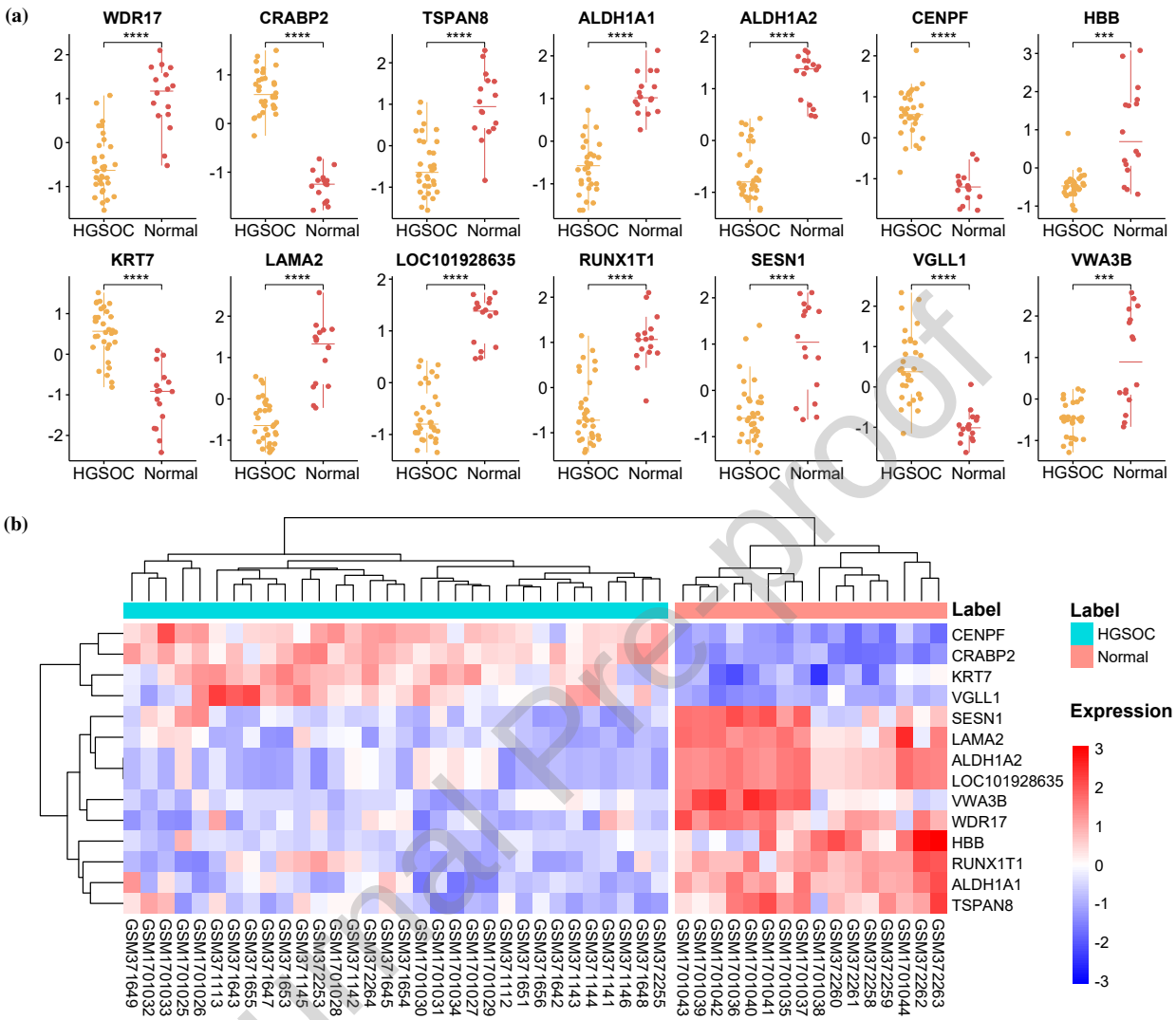


Fig. 6 The verification of 14 HGSOC biomarkers. (a) The gene expression profile for each HGSOC biomarker on the discovery data. The significance is calculated by Welch's t-test, where '*': $p < 0.05$, '**': $p < 0.01$, '***': $p < 0.001$ and '****': $p < 0.0001$. (b) The heatmap with cluster analysis.

Table 3
The classification AUC on external validations by StabML-RFE and other five comparable ML-RFE methods.

Datasets	AB-RFE	GBDT-RFE	NNET-RFE	SVM-RFE	XGB-RFE	StabML-RFE
GSE120196	0.850	0.900	0.825	0.875	0.800	**0.950**
GSE14407	0.986	**1.000**	0.972	**1.000**	**1.000**	**1.000**
GSE26712	0.903	0.965	0.978	0.984	0.949	**0.981**
GSE40595	**1.000**	0.977	0.988	**1.000**	0.977	0.992
Mean ± Std	0.935±0.061	0.960±0.037	0.941±0.067	0.965±0.052	0.931±0.079	**0.981±0.019**
# of features	Top 14	Top 14	Top 14	Top 14	Top 14	14

Std: Standard deviation.

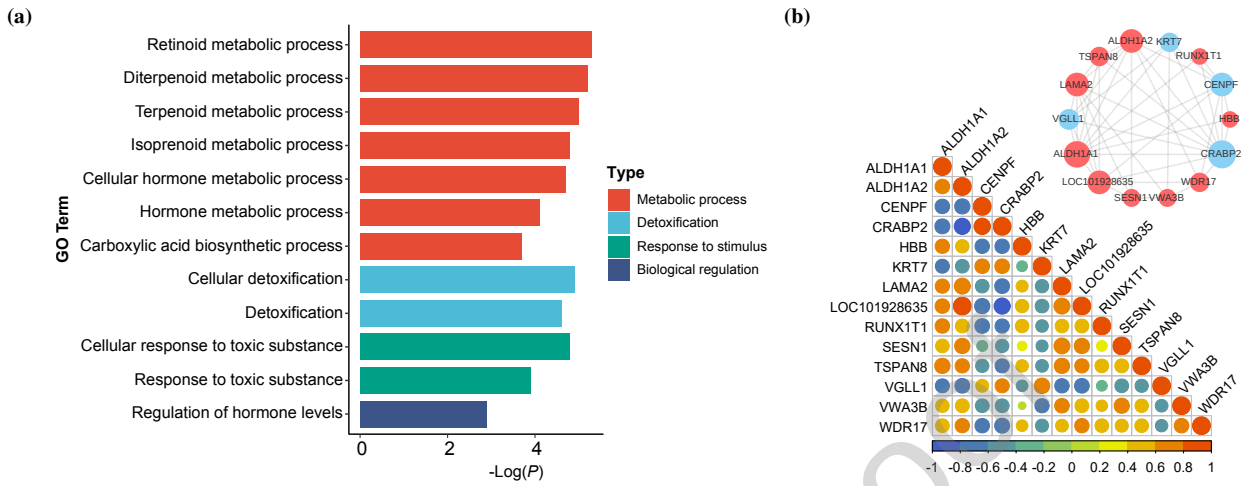


Fig. 7 The verification of 14 HGSOC biomarkers. (a) The GO enrichment analysis. (b) The correlation analysis and co-expression network, where the red nodes represent up-regulated genes and the blue nodes represent down-regulated ones.

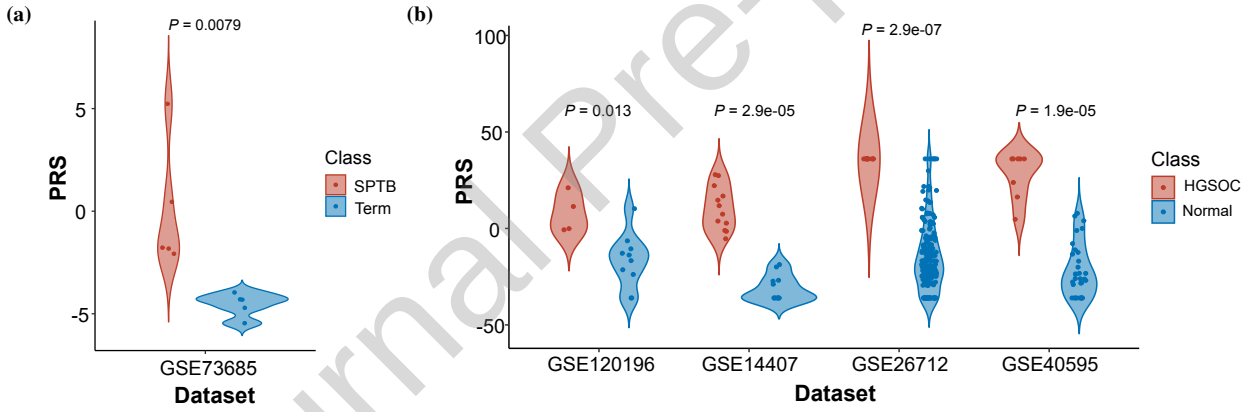


Fig. 8 The violin plot of PRS on the independent data, where the *P*-values are obtained by Wilcoxon test. (a) The PRS of SPTB and Term samples on GSE73685. (b) The PRS of HGSOC and Normal samples on GSE120196, GSE14407, GSE26712 and GSE40595.

Methods	AB-RFE	DT-RFE	GBDT-RFE	NB-RFE	NNET-RFE	RF-RFE	SVM-RFE	XGB-RFE	StabML-RFE
AB-RFE	1.000	*	0.292	*	0.348	*	*	*	0.528
DT-RFE	~	1.000	*	*	*	*	*	*	*
GBDT-RFE	0.000	~	1.000	*	0.348	*	*	*	0.528
NB-RFE	~	~	~	1.000	*	*	*	*	*
NNET-RFE	0.061	~	0.020	~	1.000	*	*	*	0.618
RF-RFE	~	~	~	~	~	1.000	*	*	*
SVM-RFE	0.061	~	0.061	~	0.209	~	1.000	*	*
XGB-RFE	0.040	~	0.000	~	0.020	~	0.040	1.000	*
StabML-RFE	0.111	~	0.081	~	0.333	~	0.429	0.081	1.000

Fig. 9 The robustness of biomarkers screened by StabML-RFE method and other ML-RFE methods, where the notations ∗s and ∼s refer to the non-existed comparable objects for SPTB and HGSOC respectively.

because of their underperforming AUC values on the internal validation. Thereafter, the XGB-RFE method is subsequently deleted based on the guide of the combination with the highest stability value. Therefore, in the upper triangular location, only the optimal features obtained from AB-RFE, DT-RFE and NNET-RFE are compared with the screened biomarkers of StabML-RFE. Clearly, the stability value of StabML-RFE is significantly higher than others. For HGSOC, the two methods, DT-RFE and NB-RFE, get lower AUC values in the internal validation. Thus they are removed firstly. Then RF-RFE method is secondly removed because it does not contribute to the stability evaluation of the optimal combination in the stable feature aggregation process. At last, AB-RFE, GBDT-RFE, NNET-RFE, SVM-RFE and XGB-RFE are left to compare their optimal features with the selected 14 biomarkers by StabML-RFE. As shown in the down triangular location, the features of StabML-RFE win bigger stability values in contrast with the other feature sets.

4. Conclusions

The gene expression data deposited in the public database provides us valuable resources for developing a computational strategy for biomarker discovery. The stability of feature selection is extremely important for biomarker discovery. In this paper, we proposed a computational pipeline named StabML-RFE for identifying robust biomarkers from gene expression data. StabML-RFE is promising that uses the stability metric as a screening evaluation for the feature genes obtained from different ML-RFE methods. In this work, we identified the robust genes based on the principle of maximizing the stability value of different optimal feature subset combinations. For justifying our findings, we performed the differential expression analysis, clustering analysis and correlation analysis. In addition to carrying out the necessary literature check, we also verified the identified biomarkers on some independent datasets for sample classification. The results clearly indicate the biomarkers are effective and efficient. The robust biomarkers of SPTB or HGSOC revealed by our proposed StabML-RFE method are great potential in providing quantitative reference and decision-making basis for early diagnosis. It provides the possibility of non-invasive alternative for the detection of SPTB and HGSOC. However, the limitation of this study is that the numbers of disease and control samples in the two diseases are still small. If we obtain more paired normal and disease samples, we can further improve the precision of discovering feature genes as diagnostic biomarkers.

CRediT authorship contribution statement

LL: Conceptualization, Methodology, Data processing, Formal analysis, Software, Writing - review & editing. WKC: Methodology, Supervision, Writing - review & editing. ZPL: Investigation, Conceptualization, Methodology, Supervision, Funding acquisition, Writing - review & editing.

Supplementary Materials

All supplementary materials can be found at `https://github.com/zpliulab/StabML-RFE`.

Methodology. The mathematical model, detailed derivation and parameter choice of the eight machine learning classifiers.

Table S1. The details of DEGs selected by EB method and adjusted by BH method.

Table S2. The functional enrichment analysis results of SPTB.

Table S3. The one-to-one correspondences and combinations of the optimal feature subsets about HGSOC.

Table S4. The GO term results of co-expression biological network using the NOA method.

Table S5. The information of 14 screened HGSOC biomarkers reported by the existing literature.

Declaration of Competing Interest

The authors declare that they have no competing interests in this work.

Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions which greatly improve our paper. The authors also thank the members in our lab at Shandong University and AMACL lab at The University of Hong Kong for their assistance in the project. This work

was partially supported by the National Natural Science Foundation of China (NSFC) under grant number 61973190; National Key Research and Development Program of China under grant number 2020YFA0712402; Key Research and Development Project of Shandong Province, China under grant number 2018GSF118043; Natural Science Foundation of Shandong Province of China (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 Program of Qilu Young Scholars of Shandong University; the Scholarship under Shandong University's Exchange Program.

References

T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (2010) 392–398. Doi: 10.1093/bioinformatics/btp630.

H. Li, Y. Guan, Machine learning empowers phosphoproteome prediction in cancers, Bioinformatics 36 (2020) 859–864. Doi: 10.1093/bioinformatics/btz639.

Z.-P. Liu, Identifying network-based biomarkers of complex diseases from high-throughput data, Biomark Med 10 (2016) 633–650. Doi: 10.2217/bmm-2015-0035.

I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach Learn 46 (2002) 389–422. Doi: 10.1023/A:1012487302797.

Y. Freund, R. E. Schapire, et al., Experiments with a new boosting algorithm, in: Procof International Conference on Machine Learning, volume 96, Citeseer, 1996, pp. 148–156. Doi: 10.1023/A:1010933404324.

J. R. Quinlan, Induction of decision trees, Mach Learn 1 (1986) 81–106. Doi: 10.1007/BF00116251.

J. H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232. Doi: 10.1214/aos/1013203451.

P. Domingos, M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Mach Learn 29 (1997) 103–130. Doi: 10.1023/A:1007413511361.

R. Hecht-Nielsen, Theory of the backpropagation neural network, in: Neural Networks for Perception, Elsevier, 1992, pp. 65–93. Doi: 10.1016/B978-0-12-741252-8.50010-8.

L. Breiman, Random forests, Mach Learn 45 (2001) 5–32.

C.-Y. Sun, T.-F. Su, N. Li, B. Zhou, E.-S. Guo, Z.-Y. Yang, J. Liao, D. Ding, Q. Xu, H. Lu, et al., A chemotherapy response classifier based on support vector machines for high-grade serous ovarian carcinoma, Oncotarget 7 (2016) 3245–3254. Doi: 10.18632/oncotarget.6569.

T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. Doi: 10.1145/2939672.2939785.

A. Bommert, T. Welchowski, M. Schmid, J. Rahnenführer, Benchmark of filter methods for feature selection in high-dimensional gene expression survival data, Brief Bioinform 23 (2022) bbab354. Doi: 10.1093/bib/bbab354.

Y. Ge, Z. He, Y. Xiang, D. Wang, Y. Yang, J. Qiu, Y. Zhou, The identification of key genes in nasopharyngeal carcinoma by bioinformatics analysis of high-throughput data, Mol. Biol. Rep. 46 (2019) 2829–2840. Doi: 10.1097/PAS.0b013e318212ae22.

D. Guan, W. Yuan, Y.-K. Lee, K. Najeebullah, M. K. Rasel, A review of ensemble learning based feature selection, IETE Technical Review 31 (2014) 190–198. Doi: 10.1080/02564602.2014.906859.

C.-W. Chen, Y.-H. Tsai, F.-R. Chang, W.-C. Lin, Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results, Expert Syst 37 (2020) e12553. Doi: 10.1111/exsy.12553.

M. Mera-Gaona, D. M. López, R. Vargas-Canas, U. Neumann, Framework for the ensemble of feature selection methods, Applied Sciences 11 (2021) 8122. Doi: 10.3390/app11178122.

W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, A. Napolitano, A review of the stability of feature selection techniques for bioinformatics data, in: 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), IEEE, 2012, pp. 356–363. Doi: 10.1109/IRI.2012.6303031.

A. Ben Brahim, M. Limam, Ensemble feature selection for high dimensional data: a new method and a comparative study, Adv Data Anal Classi 12 (2018) 937–952. Doi: 10.1007/s11634-017-0285-y.

K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, W. K. Tiong, A new hybrid ensemble feature selection framework for machine learning-based phishing detection system, Inform Sciences 484 (2019) 153–166. Doi: 10.1016/j.ins.2019.01.064.

Z. He, W. Yu, Stable feature selection for biomarker discovery, Comput Biol Chem 34 (2010) 215–225. Doi: 10.1016/j.compbiolchem.2010.07.002.

F. Farinella, M. Merone, L. Bacco, A. Capirchio, M. Ciccozzi, D. Caligiore, Machine learning analysis of high-grade serous ovarian cancer proteomic dataset reveals novel candidate biomarkers, Sci. Rep. 12 (2022) 1–12. Doi: 10.1038/s41598-022-06788-2.

S. Hwangbo, S. I. Kim, J.-H. Kim, K. J. Eoh, C. Lee, Y. T. Kim, D.-S. Suh, T. Park, Y. S. Song, Development of machine learning models to predict platinum sensitivity of high-grade serous ovarian carcinoma, Cancers 13 (2021) 1875. Doi: 10.3390/cancers13081875.

W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics 8 (2007) 118–127. Doi: 10.1093/biostatistics/kxj037.

Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of The Royal Statistical Society: Series B (Methodological) 57 (1995) 289–300. Doi: stable/2346101.

Y. J. Heng, C. E. Pennell, S. W. McDonald, A. E. Vinturache, J. Xu, M. W. Lee, L. Briollais, A. W. Lyon, D. M. Slater, A. D. Bocking, et al., Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women, PloS One 11 (2016) e0155191. Doi: 10.1371/journal.pone.0155191.

R. Bukowski, Y. Sadovsky, H. Goodarzi, H. Zhang, J. R. Biggio, M. Varner, S. Parry, F. Xiao, S. M. Esplin, W. Andrews, et al., Onset of human preterm and term birth is related to unique inflammatory transcriptome profiles at the maternal fetal interface, PeerJ 5 (2017) e3685. Doi: 10.7717/peerj.3685.

Y. Yamamoto, G. Ning, B. E. Howitt, K. Mehra, L. Wu, X. Wang, Y. Hong, F. Kern, T. S. Wei, T. Zhang, et al., In vitro and in vivo correlates of physiological and neoplastic human fallopian tube stem cells, The Journal of Pathology 238 (2016) 519–530. Doi: 10.1002/path.4649.

E. R. King, C. S. Tung, Y. T. Tsang, Z. Zu, G. T. Lok, M. T. Deavers, A. Malpica, J. K. Wolf, K. H. Lu, M. J. Birrer, et al., The anterior gradient homolog 3 (agr3) gene is associated with differentiation and survival in ovarian cancer, The American Journal of Surgical Pathology 35 (2011) 904. Doi: 10.1097/PAS.0b013e318212ae22.

C.-L. Au-Yeung, T.-L. Yeung, A. Achreja, H. Zhao, K.-P. Yip, S.-Y. Kwan, M. Onstad, J. Sheng, Y. Zhu, D. L. Baluya, et al., Itln1 modulates invasive potential and metabolic reprogramming of ovarian cancer cells in omental microenvironment, Nat Commun 11 (2020) 1–16. Doi: 10.1038/s41467-020-17383-2.

N. J. Bowen, L. Walker, L. V. Matyunina, S. Logani, K. A. Totten, B. B. Benigno, J. F. McDonald, Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells, BMC Med. Genomics 2 (2009) 1–14. Doi: 10.1186/1755-8794-2-71.

V. Vathipadiekal, V. Wang, W. Wei, L. Waldron, R. Drapkin, M. Gillette, S. Skates, M. Birrer, Creation of a human secretome: a novel composite library of human secreted proteins: validation using ovarian cancer gene expression data and a virtual secretome array, Clin Cancer Res 21 (2015) 4960–4969. Doi: 10.1158/1078-0432.CCR-14-3173.

T.-L. Yeung, C. S. Leung, K.-K. Wong, G. Samimi, M. S. Thompson, J. Liu, T. M. Zaid, S. Ghosh, M. J. Birrer, S. C. Mok, Tgf-β modulates ovarian cancer invasion by upregulating caf-derived versican in the tumor microenvironment, Cancer Res. 73 (2013) 5016–5028. Doi: 10.1158/0008-5472.CAN-13-0023.

C. Cortes, V. Vapnik, Support-vector networks, Mach Learn 20 (1995) 273–297. Doi: 10.1007/BF00994018.

Z. Zhang, Z.-P. Liu, Robust biomarker discovery for hepatocellular carcinoma from high-throughput data by multiple feature selection methods, BMC Med. Genomics 14 (2021) 1–12. Doi: 10.1186/s12920-021-00957-4.

A. Bommert, M. Lang, stabm: Stability measures for feature selection, Journal of Open Source Software 6 (2021) 3010. Doi: 10.21105/joss.03010.

K. Dunne, P. Cunningham, F. Azuaje, Solutions to instability problems with sequential wrapper-based approaches to feature selection, J Mach Learn Res 1 (2002) 22. Doi: .

L. Li, Z.-P. Liu, A connected network-regularized logistic regression model for feature selection, Applied Intelligence 52 (2022) 1–31. Doi: 10.1007/s10489-021-02877-3.

L. Li, Z.-P. Liu, Biomarker discovery for predicting spontaneous preterm birth from gene expression data by regularized logistic regression, Comput. Struct. Biotechnol. J. 18 (2020) 3434–3446. Doi: 10.1016/j.csbj.2020.10.028.

Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, S. K. Chanda, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, Nat Commun 10 (2019) 1–10. Doi: 10.1038/s41467-019-09234-6.

M. Rasmussen, M. Reddy, R. Nolan, J. Camunas-Soler, A. Khodursky, N. M. Scheller, D. E. Cantonwine, L. Engelbrechtsen, J. D. Mi, A. Dutta, et al., Rna profiles reveal signatures of future health and disease in pregnancy, Nature (2022) 1–6. Doi: 10.1038/s41586-021-04249-w.

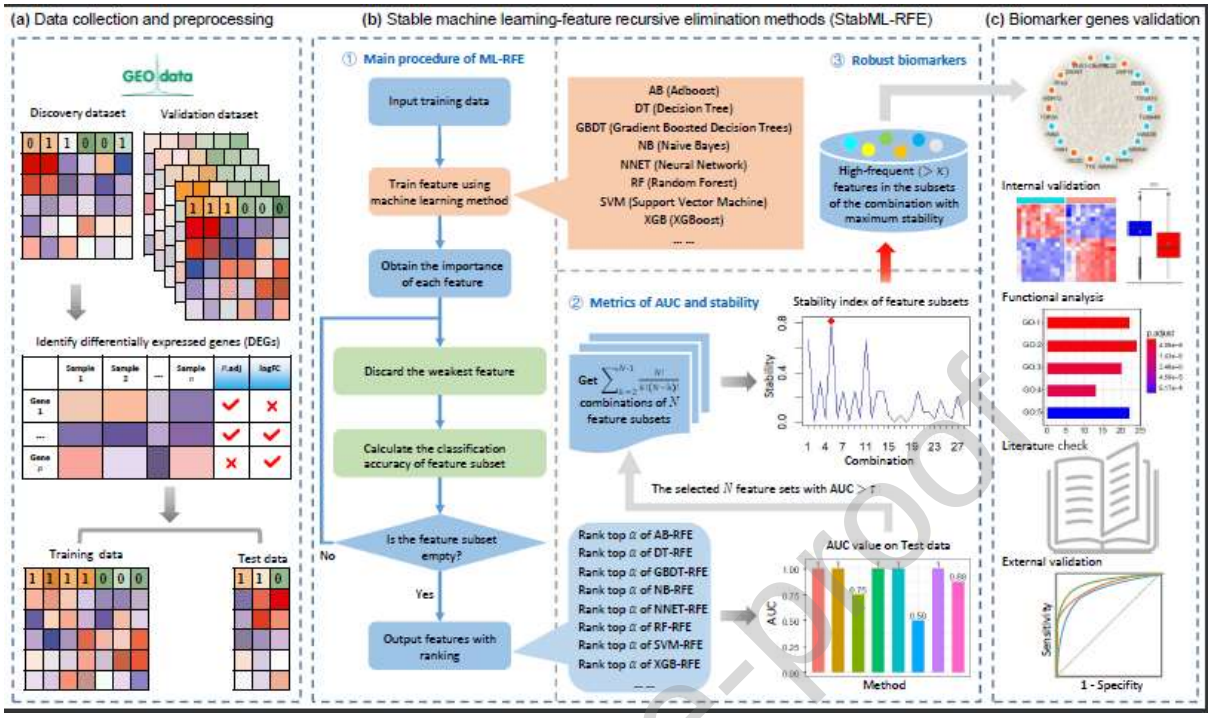
J. Plunkett, S. Doniger, T. Morgan, R. Haataja, M. Hallman, H. Puttonen, R. Menon, E. Kuczynski, E. Norwitz, V. Snegovskikh, et al., Primate-specific evolution of noncoding element insertion into pla2g4c and human preterm birth, BMC Med. Genomics 3 (2010) 1–9. Doi: 10.1186/1755-8794-3-62.

K. Yamada, K. Matsubara, Y. Matsubara, A. Watanabe, S. Kawakami, F. Ochi, K. Kuwabara, Y. Mushimoto, H. Kobayashi, Y. Hasegawa, et al., Clinical course in a patient with myopathic vlcad deficiency during pregnancy with an affected baby, JIMD Reports 49 (2019) 17–20. Doi: 10.1002/jmd2.12061.

A. Nath, P. A. Cosgrove, H. Mirsafian, E. L. Christie, L. Pflieger, B. Copeland, S. Majumdar, M. C. Cristea, E. S. Han, S. J. Lee, et al., Evolution of core archetypal phenotypes in progressive high grade serous ovarian cancer, Nat Commun 12 (2021) 1–16. Doi: 10.1038/s41467-021-23171-3.

J. Wang, Q. Huang, Z.-P. Liu, Y. Wang, L.-Y. Wu, L. Chen, X.-S. Zhang, Noa: a novel network ontology analysis method, Nucleic acids research 39 (2011) e87–e87. Doi: 10.1093/nar/gkr251.

Graphical abstract



Author Statement

LL: Conceptualization, Methodology, Data processing, Formal analysis, Software, Writing - review \& editing.

WKC: Methodology, Supervision, Writing - review \& editing.

ZPL: Investigation, Conceptualization, Methodology, Supervision, Funding acquisition, Writing - review \& editing.

Declaration of Competing Interest

The authors declare that they have no competing interests in this research.

Highlights

- A stable feature selection method (StabML-RFE) is proposed to screen robust biomarkers.
- StabML-RFE employs some popular ML-RFE methods and integrates them into an aggregation-like framework.
- StabML-RFE ensembles multiple optimal feature subsets by aggregating AUC values and stability indices.
- The robustness of screened biomarker genes is measured by the Stability metric based on Hamming distance.